Two or three things that we know about the Kohonen algorithm

M.Cottrell[†], J.C.Fort[‡], G.Pagès[*]

† Samos/Université Paris 1

90, rue de Tolbiac, F-756345 Paris Cedex 13, France

‡ Samos et Université Nancy 1/Département de Mathématiques

F-54506 Vandœuvre-Lès-Nancy Cedex, France

∗ Samos et Université Paris 6/Laboratoire de Probabilités

F-55252 Paris Cedex 05, France

Abstract

Many theoretical papers are published about the Kohonen algorithm. It is not easy to understand what is exactly proved, because of the great variety of mathematical methods. Despite all these efforts, many problems remain without solution. In this small review paper, we intend to sum up the situation.

1 Introduction

The very popular Kohonen algorithm was originally devised by Teuvo Kohonen in 1982 [12] and [13], as a model of the self-organization of neural connections. It is well known that in many cases nearby stimuli are coded on nearby cortical areas and the Kohonen algorithm is able to construct such mappings. The construction is progressive, as soon as the inputs are randomly presented, the weights are updated to reinforce the proximity between the input distribution and the discrete configuration of the weights. There is self-organization because, as far as possible, neighbors in the input space are mapped onto neighboring units. So the algorithm leads to an organized representation of the input space, from an initial complete desorder. As a matter of fact, the adaptation of the weights can be decomposed into two phases. When everything works well, one can observe self-organization (or ordering) first, and after that, convergence of the weights in order to better quantify the input space. For this second phase, the adaptation parameter has to decrease to 0.

These properties have been used in numerous practical applications that it is impossible to enumerate here. Let us only mention speech recognition (Kohonen, 1989, [15]), robotics (Ritter et al., 1989, [24]), computer vision (Oja, 1992, [20]), combinatorial optimization (Fort, 1988, [10]), classification (Kohonen, 1984, [14]), and so on. In most cases, the main feature is the self-organization property. But for other applications like automatic meshing (Sarzeaud et al., 1990, [28], or numerical integration (Pagès, 1993, [21]), the convergence property and the asymptotic values are the most worthwhile.

Despite the large use and the different implementations in multi-dimensional settings, the Kohonen algorithm is surprisingly resistant to a complete mathematical study. As far as we know, the only case where a thorough analysis could be achieved is the *one dimensional case* (the input space has dimension 1) for a *linear network* (the units are disposed along a one-dimensional array). A sketch of the proof can be encountered in the original Kohonen papers in 1982 and in his book in 1984, [12], [13], [14]. The first complete proof of the self-organization property and of the convergence was established (for uniform distribution of the inputs and a step-neighborhood function) by Cottrell and Fort in 1987, [4]. Then, these results were generalized to a very large class of input distribution by Bouton and Pagès in 1992 and 1993, [1], [2], [3]. Other papers deal with a more general form of the neighborhood function : in this framework, Erwin et al. (1992) have sketched the extension of the proof of self-organization [8] and studied the influence of the neighborhood function [7], and Fort and Pagès in 1993 [11] have rigorously proved the convergence almost sure, after self-organization.

For higher dimensions, the results are only partial. Two main difficulties can explain this situation. First, until now, nobody has been able to define what is a *correctly ordered configuration in dimension greater than 1* which would be stable for the algorithm : the grid configurations that Lo et al. proposed in 1991 or 1993, [18], [19] are in fact not stable. Secondly, Erwin et al. in 1992 [8] have proved that it is *impossible to associate a global decreasing potential function* to the algorithm, as long as the probability distribution of the inputs is continuous. They extend the first results of Tolat et al. in 1990, [30], to propose a set of single-unit energy functions, which give some description of the individual behavior. Before that, Ritter et al. in 1986 and 1988, [22], [23] have thrown some light on the stationary state in any dimension, but they study only the final phase *after the self-organization*, and do not prove the existence of a stationary state. Recently, in 1993, Fort and Pagès [11] give some results in higher dimension, but these results are very partial.

On the other hand, some theoretical results are available in different contexts. In 1992, Thiran and Hasler [29] give a proof of self-organization in dimension 1 for a large class of neighborhood function, when the input signals and the weights are *quantized* in discrete values. In 1991, Ritter et al. [25], introduce a potential function *when the input signal can only take values from a finite discrete set*, so that the Kohonen algorithm is a stochastic gradient descent on this function, at least for each constant value of the neighborhood function. For this discrete case, in 1993, Růžička [27] shows that this potential function is not differentiable, but gives conditions on the adaptation parameter and the neighborhood function for which the convergence of the algorithm to a stationary point is proved. At last, in 1992, Pagès [21] completely studies the so-called *0 neighbor algorithm* which always corresponds to a gradient descent procedure.

2 The Kohonen algorithm

The network has one layer of n units arranged in a lattice, (generally in a one- or two-dimensional array). The set $I = \{1, \ldots, , n\}$ of the units is endowed with a topological structure provided by a neighborhood function Λ defined on $I \times I$. We consider a symmetrical and decreasing neighborhood function, such that $\Lambda(i, j) = \Lambda(j, i)$ depends only on the distance between i and j in the array, ($\mid i - j \mid$ if the array is one-dimensional. $\Lambda(i, j)$ decreases with increasing distance between i and j, and $\Lambda(i,i)$ is usually equal to 1.

The input space Ω is included in $\mathcal{R}^d$. The units are fully connected to the inputs, and X_{ij} represents the strengh of connection between unit i and the j^{th} component of the input.

The trick consists in representing the unit i by the vector

$$X_i = (X_{i1},\, X_{i2},\, \ldots,\, X_{id}).$$

By normalizing, if necessary, the vectors $X_1, X_2, \ldots, X_n$, we can represent them in the same space Ω as inputs. Then, the network state at time t is given by

$$X(t) = (X_1(t),\, X_2(t),\, \ldots,\, X_n(t)).$$

For a given state X, the network response to input ω is the **winner** unit i_0, the closest unit to input ω. Thus, the network defines a map $\Phi_X : \omega \longmapsto i(\omega, X)$, from Ω to $I = \{1, \ldots, n\}$, and the goal of the learning algorithm is to converge to a network state such as the corresponding map will be topology preserving. For a given state X, let us denote $C_i(X)$ the set of the inputs such that i is the winner unit for them, that is $C_i(X) = \Phi_X^{-1}(i)$. The set of the classes $C_i(X)$ is the Euclidean Voronoï tesselation of the space Ω related to X.

The algorithm is as follows :

- at time $t = 0$, X_i^0 is chosen at random,
- if $X(t)$ is the current state,
  - present input $\omega(t+1)$ chosen in Ω according to a distribution $\mathcal{P}$,
  - compute the best matching unit i_0 by

$$\text{dist}(X_{i_0}(t),\, \omega(t+1)) = \min_j \text{dist}(X_j(t),\, \omega(t+1))$$

  - update the weights by

$$X_i(t+1) = X_i(t) - \varepsilon_t \Lambda(i_0, i)(X_i(t) - \omega(t+1))$$

    for each $i \in I$.

This rule strengthens the similarity between the input $\omega(t+1)$ and the responses of unit i_0 and of its neighbours.

The essential parameters are the adaptation parameter ε_t, which is “small” and positive, constant or decreasing with time, the neighborhood function Λ, which can be constant or time dependent, the dimension d of the input space, the probability distribution $\mathcal{P}$.

As the inputs are i.i.d. random variables with probability distribution $\mathcal{P}$ on Ω, the network state at time t is a random Ω^n-valued vector $X(t)$ displaying as :

$$X(t+1) = X(t) - \varepsilon_t\, H(\omega(t+1),\, X(t)) \tag{1}$$

Notice that according to (1), $X(t)$ is always a special case of stochastic process, that is a *Markov chain*, which is homogeneous in time if and only if ε_t and Λ are time-invariant.

3 Description of the mathematical tools

As we observe before, one has to separate two kinds of results : those related to the self-organization, from those related to convergence after organization. In both cases, almost all the results have been obtained for a time-invariant neighborhood function.

To analyse the *self-organization*, one uses mainly *Markov chain techniques.* One tries to prove that there exists some *absorbing set* (which could be qualified as *ordered*), which the process $X(t)$ enters with probability 1 after a finite time, [4], [1], [8]. A very useful way to find these absorbing sets can be to put in evidence a function which decreases along each trajectory of the process [9], [6]. This function can be viewed as an *energy function*, whose minima correspond to what we call *ordered states* and are the absorbing sets.

As to the *convergence phase*, the techniques depend on the kind of desired convergence. For the *almost sure* convergence, the parameter ε_t has to decrease to 0, and the form of the equation (1) suggests to consider the Kohonen algorithm as a Robbins-Monro [26] algorithm. The usual hypothesis on the adaptation parameter is then :

$$\sum_t \varepsilon_t = +\infty \text{ and } \sum_t \varepsilon_t^2 < +\infty \tag{2}$$

One can observe that all the possible limit states $x^\star$ are solutions of

$$h(x) = 0$$

where $h(x) = E(H(\omega, x) = \int H(\omega, x)\, d\mathcal{P}(\omega)$ is the expectation of $H(., x)$ with respect to the probability measure $\mathcal{P}$. Some *global assumptions* on the function h (and on its *gradient*) are needed to ensure the almost sure convergence. If only some *local assumptions* on the function h are available, one can follow the "Kushner and Clark" way which is based on the study of the equilibria of the average ordinary differential equation (O.D.E.) [17]

$$\frac{dx}{dt} = -\ h(x)$$

Many problems are encountered : how to compute these equilibria, are they unique, are they attractive ? In the best of cases, it is then possible to get a weakened convergence (called $K\&C$ convergence), which consists in ensuring the almost sure convergence to some equilibrium $x^\star$, as long as the process $X(t)$ comes back infinitely often in some attracting area of $x^\star$.

Another method consists in trying to put in evidence that the process defined by (1) can be viewed as a *stochastic gradient descent* associated with some potential function, [30], [21], [25], [8]. In that case, if (2) is verified, it is sure that the process converges towards a zero of h. But nothing ensures automatically that these zeros are minima, and even if they are minima, they do not necessarily correspond to *ordered situations*, [6]. Moreover, it also happens that the potential function can be differentiable not everywhere.

If the *convergence in distribution* is desired, that is the establishment of some invariant probability measure, (and not the stabilization in some fixed values), the adaptation parameter has to remain constant and the relevant techniques are Markovian.

In spite of the great variety of mathematical tools and unfortunately, the set of known results is very small with respect to the set of possible answers. In the following sections, we try to sum up the theoretical available results, in each main case.

4 The self-organization for dimension 1

In this case, the input space is $[0,1]$ (the dimension d is 1) and the units are arranged along a linear array. The neighborhood function Λ is supposed to be decreasing with the distance between units. The stimuli distribution is continuous on $[0,1]$. This means that it does not charge any point, and it is true for any distribution having a density. Let us define $F_n^+ = \{x \in \mathcal{R} \,/\, 0 < x_1 < x_2 < \ldots < x_n < 1\}$ and $F_n^- = \{x \in \mathcal{R} \,/\, 0 < x_n < x_{n-1} < \ldots < x_1 < 1\}$. One can see that the number of inversions (badly ordered 3-uples) is a non increasing function associated to the algorithm, which is 0 on $F_n^+ \cup F_n^-$.

In [4], [1], [11], the following results are proved :

Theorem 1 *(a) The two sets F_n^+ and F_n^- are absorbing sets.*
(b) If ε is a constant, and if Λ is the step function defined by $\Lambda(i,j) = 1$ iff $\mid i - j \mid \leq 1$, the re-ordering time τ, that is the hitting time in $F_n^+ \cup F_n^-$, is almost surely finite, and $\exists \lambda > 0$, s.t. $\sup_{x \in [0,1]^n} E_x(\exp(\lambda\tau)) < +\infty$, where E_x is the expectation given $X(0) = x$.

Remarks : i) The inequality gives an upper bound of the re-ordering time.
ii) In [8], the arguments to extend the part (b) to the general decreasing neighborhood function case are sketched.
iii) With the same arguments as in [2], it is possible to prove that if $\sum \varepsilon_t = +\infty$, then $\forall x \in [0,1]^n$, Proba $(\tau < \infty)$ is positive, but not sure.
iv) The hypothesis on the distribution $\mathcal{P}$ (continuity) seems minimal : in fact, if $\mathcal{P}$ is uniform with two values a and b, $0 < a < b < 1$, and $n = 4$, there is no re-ordering if the initial value $X(0) = x$ satisfies $0 < a < x_2 < x_1 < x_3 < x_4 < b < 1$.

5 The convergence for dimension 1

In [4] for the uniform distribution, [1] and more recently [11], the almost sure convergence is proved, according to the next theorem :

Theorem 2 *Let us assume that ε satisfies the condition (2), and that the neighborhood function is decreasing enough. Assume also that the support of the probability $\mathcal{P}$ is $[0,1]$ (i.e. there is no closed subset in $[0,1]$ with probability 1).*
Then
(a) The mean function h has at least a zero $x^\star$ in F_n^+
(b) Let the input distribution $\mathcal{P}$ have a density f such that $f > 0$ on $]0,1[$ and let $\ln(f)$ be strictly concave (or $\ln(f)$ be only concave, with $\lim_{0^+} f + \lim_{1^-} f$ strictly positive), then every stationary point $x^\star$ is attracting. So if $X(0) \in F_n^+$, $X(t) \xrightarrow{a.s} x^\star$ in the Kushner & Clark sense.

(c) If the input distribution is uniform on $[0,1]$, $x^\star$ *is unique and* $X(t) \xrightarrow{a.s} x^\star$, *if* $X(0) \in F_n^+$.

The exact condition for Λ is : $(n \geq 2$ and $\Lambda(i,j) < 1, for \mid i-j \mid= 1)$ or $(n \geq 3$ and $\Lambda(i,j) < 1, for \mid i-j \mid= 2)$ or $(n \geq 5$ and $\Lambda(i,j) < 1, for \mid i-j \mid= 3)$.

Remarks : i) Almost all the usual probability distributions (truncated on $[0,1]$) have a density such that $\ln(f)$ is concave

ii) The unicity of the equilibrium (in F_n^+) is proved only for the uniform distribution

iii) The same properties hold for F_n^-

iv) The condition (2) is essential, because if ε_t decreasing and $\sum_t \varepsilon_t = +\infty$, there is only convergence in probability

v) It is possible to build counter-examples with no ln-concave density and where convergence to the stationary point does not hold. (See section 6).

In [4] and [2], the convergence in distribution is proved, for a *constant* ε, a *two-neighbor step function*, and assuming some condition on the distribution $\mathcal{P}$ (it has a lower bounded density at least on a small open set in $[0,1]$).

6 The 0 neighbor case

This case is characterized by choosing the neighborhood function $\Lambda(i,j) = 1$ if $i = j$, and 0 elsewhere. There is no topology on I, and no *reordering*. This algorithm is known as a "space quantization algorithm". The dimension d of the input space can be greater than 1.

The main result is :

i) The 0-neighbor algorithm derives from the potential

$$V(x) = \int \min_i \|x_i - \omega\|^2 d\mathcal{P}(\omega)$$

ii) If the distribution probability $\mathcal{P}$ is continuous, or for example has a density f,

$$V(x) = \sum_{i=1}^{n} \int_{C_i(x)} \|x_i - \omega\|^2 f(\omega) d\omega$$

V can be interpreted as an intra-class variance. But in fact, this potential function is not everywhere differentiable and the complete analysis is not easy [24], [21]. There are many quilibria [24], and some of them can be repulsive [21]. There is no hope to state a global almost sure convergence, except in the uniformly distributed case. In this case, one has :

Theorem 3 *If* $d = 1$ *and the stimuli are uniformly distributed on* $[0,1]$, *then :*

$$X(t) \xrightarrow{a.s.} x^\star = ((2i-1)/2n)_{1 \leq i \leq n}$$

In the general case if the density is ln-concave, with $f(0_+) + f(1_-) > 0$, (or strictly ln-concave, cf Theorem 2) we have this partial statement [21],

Theorem 4 *(a) If $d = 1$, V has only strict local minima and $X(t)$ converges almost surely to one of these local minima*
(b) When $d > 1$, $X(t)$ converges, in the Kushner & Clark sense, to a zero of h
(c) For any d, let n be a variable, V_n the corresponding potential and $x_n^\star$ one minimum of V_n. Then, the empirical discrete probability measure

$$\mathcal{P}_n \ = \ \sum_{i=1}^{n} \mathcal{P}(C_i(x_n^\star))\delta_{x_n^\star}$$

converges in distribution to the probability measure $\mathcal{P}$, when $n \to \infty$.

The property (c) properly defines the *quantization property*, and explains why the 0-neighbor algorithm provides a skeleton of the input distribution.

It is easy to build a counter example when the density f is not ln-concave: let g be a symmetric probability density, such that $g(u) = g(1-u)$. Let us consider f_n : $f_n(u) = g(nu - i)$, for $u \in [i/n, (i+1)/n]$. Then $x^\star = \{(2i-1)/2n), i = 1, \ldots, n\}$ is an equilibrium. But even if $X(t)$ converges when $g(0) \leq 1$, it is not true when $g(0) \geq 2n/(n-1)$, because in that case $x^\star$ is repulsive.

The *convergence in distribution* is true, with constant adaptation parameter ε, when the distribution $\mathcal{P}$ has no hole and has a density which is lower bounded at least on a small open set.

7 The discrete case

In this case, there is a finite number N of inputs $\{\omega_1, \omega_2, \ldots, \omega_N\}$ and the input distribution is discrete and uniform. It is the more useful setting for most practical applications, like classification or data analysis.

The main result ([16], [24]) is that for constant neighborhood (or for suitably decreasing ones [27]), the algorithm derives from the potential

$$V(x) = \frac{1}{N}\sum_{i=1}^{n} \sum_{\omega_l \in C_i(x)} (\sum_{j=1}^{n} \Lambda(i-j)\|x_j - \omega_l\|^2)$$

V is an intra-class variance extended to the neighbor classes. But this potential is not a smooth function and its study is lenghthy. The complete analysis is not achieved, even if the discrete algorithm can be viewed as a stochastic gradient descent procedure. The most complete results can be found in [27].

8 The multidimensional extension

When the dimension d is greater than 1, the available results are very few. The main reason seems to be the fact that no absorbing sets have been found. The configurations which are monotoneous in each coordinate are not stable, contrary to the intuition. Some people think that the Kohonen algorithm in dimension greater than 1 could correspond to an irreducible chain, that is a chain for which there always exists a path with positive probability to go from anywhere to everywhere. That property would imply that there is no absorbing set at all. Actually, as soon as $d \geq 2$, the O-neighbor algorithm is an irreducible chain.

Let us consider $I = I_1 \times I_2 \times \ldots \times I_d$ a d-dimensional array, with $I_l = \{1, 2, \ldots, n_l\}$, for $1 \leq l \leq d$. Let us assume that the neighborhood function is a product function (for example eight neighbors for $d = 2$) and that the input distributions in each coordinate are independent, that is $\mathcal{P} = \mathcal{P}_1 \bigotimes \ldots \bigotimes \mathcal{P}_d$. Finally, let us suppose that the support of each $\mathcal{P}_l$ is [0,1].

Let us call *grid states* the states $x^\star = (x^\star_{i_l l}, 1 \leq i_l \leq n_l, 1 \leq l \leq d)$, such that for every $1 \leq l \leq d$, $(X^\star_{i_l l}, 1 \leq i_l \leq n_l)$ is an equilibrium for the one-dimensional algorithm. Then the following results hold [11] :

Theorem 5 *(a) The grid states are equilibria of the d-dimensional algorithm. (b) For $d = 2$ and 0 neighbor, the grid equilibria*

$$x^\star = ((2i_1 - 1)/2n_1,\ (2i_2 - 1)2n_2)$$

, for $i_1 \in I_1, i_2 \in I_2$, with $n = n_1\, n_2$, and $\mathcal{P}$ the Uniform distribution $U([0,1]^2)$ is **never stable** *when $n_1/n_2 \notin [\sqrt{3}/3, \sqrt{3}]$.*

That is all for the moment. In [11], the gradient of h is computed in the d-dimensional setting. In [23], the convergence is studied after re-ordering, assuming that the organization is over, although there is no proof of it.

9 Practical and provisional remarks

In spite of, or because of the mysteries which remain to be understood, the practicians have developed a great know-how. For example, it is well-known that to accelerate the convergence of the algorithm, the adaptation parameter and the neighborhood have to be large at the beginning. On the other hand, when ε decreases to 0, it is possible to get stuck in a "metastable state". In order to avoid it, it is very useful to have at our disposal some criterium which approximately indicates whether the topology preserving property holds. See for example [5] or [31] which propose such indices. It would be probably interesting also to study the variations of the function $V(x)$, even if it is not a true potential function.

Two perspectives are promising : to carry on the work without respite as a challenge, or to define other self-organization algorithms which would be more tender.

References

[1] C.Bouton, G.Pagès, "Self-organization of the one-dimensional Kohonen algorithm" and "Convergence (a.s. and in distribution) of the one-dimensional Kohonen algorithm", Proceedings of *Aspects théoriques des réseaux de neurones*, Congrès Satellite du Congrès Européen de Mathématiques, Paris, 2-3 Juillet 1992.

[2] C.Bouton, G.Pagès, "Self-organization of the one-dimensional Kohonen algorithm with non-uniformly distributed stimuli", *Stochastic Processes and their Applications*, 47, 249-274, 1993.

[3] C.Bouton, G.Pagès, "Convergence in distribution of the one-dimensional Kohonen algorithm when the stimuli are not uniform", to appear in *Advanced in Applied Probability*, 26, 1, March 1994.

[4] M.Cottrell, J.C.Fort, "Etude d'un algorithme d'auto-organisation", *Ann. Inst. Henri Poincaré*, 23, 1, 1-20, 1987.

[5] P.Demartines, "Organization measures and representations of Kohonen maps", in : J.Hérault (ed), *First IFIP Working Group 10.6 Workshop*, 1992.

[6] M.Duflo, *Méthodes récursives aléatoires*, Masson, Paris, 369 p., 1990.

[7] E.Erwin, K.Obermayer and K.Shulten, "Self-organizing maps : stationary states, metastability and convergence rate", *Biol. Cyb.*, 67, 35-45, 1992.

[8] E.Erwin, K.Obermayer and K.Shulten, "Self-organizing maps : ordering, convergence properties and energy functions", *Biol. Cyb.*, 67, 47-55, 1992.

[9] W.Feller, *An introduction to probability theory and its applications*, Vol. 1 and 2, 3^{rd} edition, Wiley, 1968 and 1971.

[10] J.C.Fort, "Solving a combinatorial problem via self-organizing process : an application of the Kohonen algorithm to the travelling salesman problem", *Biol. Cyb.*, 59, 33-40, 1988.

[11] J.C.Fort and G.Pagès, "Sur la convergence $p.s.$ de l'algorithme de Kohonen généralisé", *Note aux Comptes Rendus de l'Académie des Sciences de Paris*, t. 317, Série I, 389-394, 1993 (developed in Preprint SAMOS #29).

[12] T.Kohonen, "Self-organized formation of topologically correct feature maps", *Biol. Cyb.*, 43, 59-69, 1982.

[13] T.Kohonen, "Analysis of a simple self-organizing process", *Biol. Cyb.*, 44, 135-140, 1982.

[14] T.Kohonen, *Self-organization and associative memory*, Springer, New York Berlin Heidelberg, 1984 (3^{rd} edition 1989).

[15] T.Kohonen, "Speech recognition based on topology preserving neural maps", in : I.Aleksander (ed) *Neural Computation* Kogan Page, London, 1989.

[16] T.Kohonen, "Self-organizing maps : optimization approaches", in : T.Kohonen et al. (eds) *Artificial neural networks, vol. II*, North Holland, Amsterdam, 981-990, 1991.

[17] H.J.Kushner, D.S.Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Volume 26, in Applied Math. Science Series, Springer, 1978.

[18] Z.P.Lo, B.Bavarian, "On the rate of convergence in topology preserving neural networks", *Biol. Cyb*, 65, 55-63, 1991.

[19] Z.P.Lo, Y.Yu and B.Bavarian, "Analysis of the convergence properties of topology preserving neural networks", *IEEE trans. on Neural Networks*, 4, 2, 207-220, 1993.

[20] E.Oja, "Self-organizing maps and computer vision", in : H.Wechsler (ed), *Neural networks for Perception*, vol.1, Academic Press, Boston.

[21] G.Pagès, "Voronoï tesselation, space quantization algorithms and numerical integration", in *Proc. of the ESANN93 Conference*, Brussels, D Facto Ed., (ISBN-2-9600049-0-6), April 1993, pp. 221-228.

[22] H.Ritter and K. Schulten, "On the stationary state of Kohonen's self-organizing sensory mapping", *Biol. Cybern.*, 54, 99-106, 1986.

[23] H.Ritter and K. Schulten, "Convergence properties of Kohonen's topology conserving maps: fluctuations, stability and dimension selection", *Biol. Cybern.*, 60, 59-71, 1988.

[24] H.Ritter T.Martinetz and K. Schulten, "Topology conserving maps for motor control", *Neural Networks, from Models to Applications*", (L.Personnaz and G.Dreyfus eds.), IDSET, Paris, 1989.

[25] H.Ritter T.Martinetz and K. Schulten, *Neural computation and Self-Organizing Maps, an Introduction*, Addison-Wesley, Reading, 1992.

[26] H.Robbins and S. Monro, "A stochastic approximation method", *Ann. Math. Stat.*, vol. 22, pp. 400-407, 1951.

[27] P.Růžička, "On convergence of learning algorithm for topological maps", *Neural Network World*, 4, 413-424, 1993.

[28] O.Sarzeaud, Y.Stéphan and C.Touzet, " Application des cartes auto-organisatrices à la génération de maillages aux éléments finis", in : *Proc. of Neuro-Nîmes*, vol. 1, 81-96, 1990.

[29] P.Thiran, M.Hasler, "Quantization effects in Kohonen networks", in : *Proc. of the Congrès Satellite du Congrès Européen de mathématiques sur les Aspects Théoriques des Réseaux de Neurones*, Paris, 1992 (submitted).

[30] V.V.Tolat, "An analysis of Kohonen's self-organizing maps using a system of energy functions", *Biol. Cyb.*, 64, 155-164, 1990.

[31] S.Zrehen, F.Blayo, "A geometric organization measure for Kohonen's map", in : *Proc. of Neuro-Nîmes*, 603-610, 1992.