SOM's Mathematics

J.C. Fort
Laboratoire de Statistique et Probabilités
118 route de Narbonne
Toulouse. France
fort@cict.fr

Abstract - *Since the discovery of the SOMs by T. Kohonen, many results have been found in order to get a better description of their behaviour. Most of them are very convincing but from a mathematical point of view, only a few are actually proved. In this paper, we make a review of some results that are still to be proved and give some framework to formulate various questions.*

Key words - Magnification factor, grid equilibrium, topology preservation, limit theorems

1 Introduction

This paper is an opportunity to review the mathematical results established about the now classical Kohonen algorithm. While the applications of the Self Organizing Maps (SOM) are numerous, only a few results is available. Moreover most of them are concerned with the one-dimensionnal case which is a very particular framework far from the applications. Nevertheless a mathematical study is needed for at least two goals. The first one is to actually prove observed facts, which could lead to a better knowledge of the behaviour of the algorithm. The second one is to better understand what is the Self Organization in order to propose some possible new algorithms based on this knowledge.

This paper is self-organized as it follows: we begin with some analytical results which are well-known but require a more rigourous proof. Then we look at the convergence of the learning algorithm. We shortly describe what is proved, which is very few with respect to what is needed and used for the applications. At this stage, the question of "what is the organization" is discussed. After these somewhat theoretical aspects, we investigate some more statistical problems.

2 Analytical Results

2.1 Magnification factor

We begin with one of the most discussed topic in the framework of the quantization and more recently of the SOM: the magnification factor. This is the fact that, when the units are equally weighted, then the simplest reconstructed distribution is biaised. If we denote by $x^{(n)} = (x_1^{(n)}, ...x_n^{(n)})$ the values of the n units minimizing the square-distortion in the quantization problem (or 0-neighbour Kohonen algorithm) the simplest reconstruction of the data distribution is given by :

$$\mu^n_{x^{(n)}} := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^{(n)}}, \; n \geq 1$$

and $\mu^n_{x^{(n)}}$ weakly converges when n goes to ∞, to the distribution

$$\mu^\infty := \frac{f^{\frac{d}{d+2}}(\xi)}{\int_{\mathbb{R}^d} f^{\frac{d}{d+2}}(u)\, du} d\xi$$

if the data distribution has a *p.d.f.* f over $\mathbb{R}^d$. The magnification factor is $\alpha = \frac{d}{d+2}$. The same phenomenon occurs in the SOM case as mentionned by T. Kohonen in 1982 ([7]). Of course the more effective reconstruction which is given by (see [6])

$$\widetilde{\mu}^n_{x^{(n)}} := \sum_{1\leq i\leq n} \mu(C_i(x^{(n)}))\delta_{x_i^{(n)}}$$

avoids this problem, where $C_i(x^{(n)})$ is the Voronoï tessel of $x_i^{(n)}$. It is easily seen that $\widetilde{\mu}^n_{x^{(n)}}$ weakly converges towards $\mu = f(\xi)d\xi$. Moreover one can prove that the "on line" estimation of the μ-mass of $C_i(x^{(n)})$ *a.s.* converges to the true value.

Reconstruction of data distribution

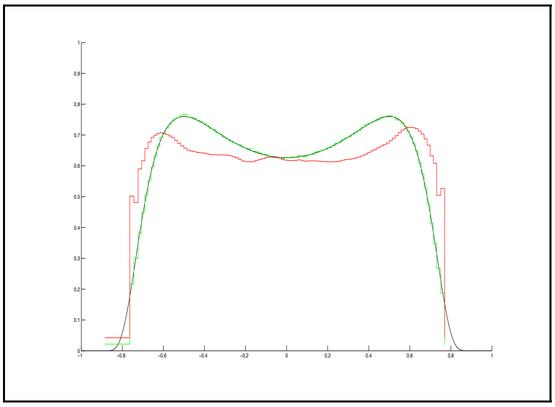

Fig 1. The *p.d.f.* reconstruction: with $\mu^n_{x^{(n)}}$ in red (light gray) and with $\widetilde{\mu}^n_{x^{(n)}}$ in green (dark gray). The true f is superposed to the red curve.

So our opinion is that this phenomenon is a purely theoretical question, but it needs some mathematical treatment to be established in the SOM case.

Ritter & Schulten ([11]), Ritter ([10]) have given an asymptotic expansion (when n goes to ∞) of α) in the 1-dimensional case with the k nearest neighbours function,

$$\alpha \sim \frac{2}{3} - \frac{1}{3(k^2+(k+1)^2)}.$$

To obtain this expansion, a very strong asumption is needed: $x_i^{(n)}$ is asymptotically a twice derivable function with respect to the "discrete" variable i.
Thus two facts are to be proved:

1. to justify the asymptotic expansion we need the existence of a smooth map g: $u \in [0,1] \longrightarrow x_u = g(u)$, which is the limiting value (for uniform convergence) of $g^{(n)}$: $\frac{i}{n} \longrightarrow x_i^{(n)}$ for $i = 1, \ldots, n$.
2. to connect $\mu^n_{x^{(n)}}$ and $\widetilde{\mu}^n_{x^{(n)}}$ we have to prove that:

$$\sup_{1 \leq i \leq n} \left| n\, \mu(C_i(x^{(n)})) - f^{1-\alpha}(x_i^{(n)}) \int_{\mathbb{R}^d} f^\alpha(u)\, du \right| \overset{n \to +\infty}{\longrightarrow} 0.$$

2.2 Grid equilibrium

The next problem is concerned with a well-known fact illustrated as follows

The grid equilibrium is stable, case $\mathbf{U}[0,1]^2$

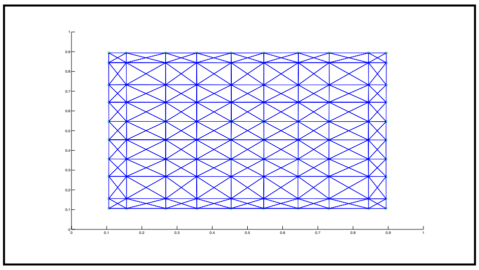

Fig 2. The equilibrium obtained for a 8 neighbours function is a product grid equilibrium.

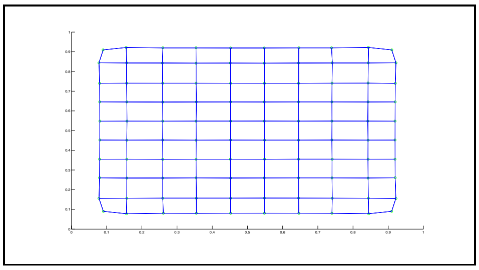

Fig 3. The equilibrium obtained for a 4 neighbours function is not a grid (see the corners).

If we denote by h the mean field of the SOM, with a neighbourhood function σ, the stability of an equilibrium point depends on the eigenvalues of the following gradient (see [5]):

$$\frac{\partial h_i}{\partial x_j} = \sum_{k \in I} \sigma(i,k) \mu(C_k(x)) \delta_{ij} e_i + \sum_{k \neq j \in I} (\sigma(i,k) - \sigma(i,j)) \int_{\bar{C}_k(x) \cap \bar{C}_j(x)} (x_i^l - \omega^l)$$

$$\times \left(\frac{1}{2} n_x^{kj} + \frac{1}{\|x_k - x_i\|} (\frac{x_k + x_j}{2} - \omega) \right) f(\omega) \lambda_x^{kj}(d\omega),$$

where $\lambda_x^{kj}(d\omega)$ is the Lebesgue measure on the border of the Voronoï tessel between x_i and x_j (the median hyperplan).
The three following results are still unproved:

1. In the case of the uniform distribution over $[0,1]^2$ with a 8-neighbours square grid, the product grid equilibrium is stable.

2. More generally in the case of uniform distribution over $[0,1]^d$ with a 2^d-neighbours "square" grid, the product grid equilibrium is stable.

3. It generalizes to the case of product of symmetric distributions.

The figures 4 and 5 give two examples of observed stable grid equilibria for product of symmetric (non uniform) distributions, when the neighbourhood function is the 8 neighbours function.

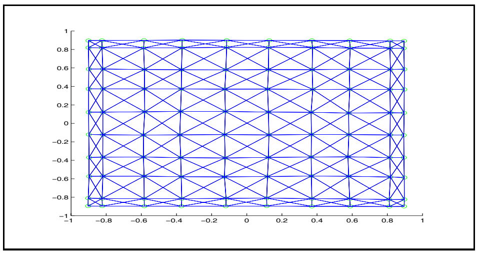

Fig 4.The equilibrium obtained for a 8 neighbours function with a product of two troncated gaussians.

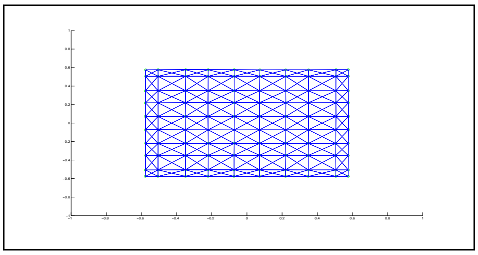

Fig 5. The equilibrium obtained for a 8 neighbours function with the product of symmetric distribution of fig 1 by itself.

2.3 Uniqueness

To end this section we mention the question of the uniqueness in dimension one. The best available result is (see [1]):

Theorem 1 *Assume that the neigbourhood function σ satisfies H_σ : for some $k_0 \leqslant \frac{n-1}{2}$, it holds $\sigma(k_{0+1}) < \sigma(k_0)$. Then if μ has a density f which is positive on $]0,1[$ and strictly Log-concave, or if f is Log-concave and $f(0_+) + f(1_-) > 0$, there exists only one equilibrium point and it is stable.*

This includes the case of uniform, gaussian, exponential distributions but not χ^2 and power distributions. However in the case of quantization, it remains true for these distributions (see [4]).
Hence the last open analytical problem we mention, is quite fuzzy: How to extend uniqueness results ?

3 Convergence of the Algorithm

We begin by the simplest case which reduces to the **one-dimensional ordering**. There are two frameworks: the constant and the decreasing step frameworks.

3.1 Constant learning step $\varepsilon_t = \varepsilon$

Theorem 2 *(see [2, 3])Let T be the ordering time, there exists $\lambda > 0$ such that:*

$$\forall x \in [0,1]^n, \quad \mathbb{E}_x e^{\lambda T} < c < +\infty.$$

Theorem 3 *(see [1]) Assume that the conditions for the existence and uniqueness of a stable equilibrium x^* are satisfied, and that the initial state is ordered. For any $0 < \varepsilon \leq \varepsilon_0 < 1$ there exists at least an invariant probability measure π^ε. Let $(\pi^\varepsilon, \varepsilon \leq \varepsilon_0)$ a family of such invariant distributions, then we have for the weak convergence of probability measures:*

$$\lim_{\varepsilon \to 0} \pi^\varepsilon = \delta_{x^*}.$$

3.2 Decreasing learning step ε_t

Theorem 4 *(see [3]) Assume the conditions for the existence and uniqueness of a stable equilibrium x^* are satisfied, then from any ordered initial state, the Kohonen algorithm converges to this equilibrium provided that $\sum \varepsilon_{t+1} = \infty$ and $\sum \varepsilon_{t+1}^2 < \infty$.*

In view of applications (the one dimensional case is of course only a test case) it would be nice to prove that:

There exists some sequence $\varepsilon_t > 0$ with limit 0, such that T is *a.s.* finite and the algorithm converges towards x^* in probability.
Obviously we have some guidelines to handle a possible proof: first we can use the existence of the Laplace transform of T to find a decreasing sequence leading to ordering, and then follow the proof of convergence of simulated annealing based on the existence of invariant measures at constant temperature concentrating on the set of global minima, to show that the algorithm concentrates on x^*.

3.3 Multi-dimensional case, constant learning step

A more difficult framework is the multidimensional case. The main result is by Sadeghi ([12]) First we assume that $\sigma(i,i) = 1$ and that for $i \neq j$, $\sigma(i,j) < 1$, and we consider a constant learning step ε. The probability distribution has a positive p.d.f. on a compact set with a non empty interior. We define the distance in variation between two probability measures on $\mathbb{R}^d$ P et Q by: $\|P - Q\| = \sup_{A \subset \mathbb{R}^d} |P(A) - Q(A)|$. Then we have the following result:

Theorem 5 *Assume that:*

- *There exists a set $I_0 \neq \emptyset$ such that there exists at least one k_0 for which $\sigma(k_0, j) = 0, \forall j \in I_0$.*
- *If $i \neq k_0$ then there exists $j \in I_0$ such that $\sigma(i, j) > 0$.*

Then the SOM algorithm X^t weakly converges, with a uniform geometric rate: let P_x^t the distribution of X^t when $X^0 = x$, there exists a unique stationnary probability distribution π^ε such that

$$\exists R_\varepsilon \in \mathbb{R}^+, \exists r_\varepsilon, 0 < r_\varepsilon < 1, \left\|\mathbf{P_x^t} - \pi^\varepsilon\right\| \leq \mathbf{R}_\varepsilon \mathbf{r}_\varepsilon^\mathbf{t}.$$

Of course this is a very general result, but it does not give any information on the nature of the stationary distribution. The tools used to prove it, come from the general theory of continuous space Markov chains. Indeed, as noticed by Bouton-Pagès ([2]), the Kohonen algorithm is not a Feller Markov chain. Yet it is a so-called T-chain as proved by Sadeghi.

Some natural further questions are:

1. Which cases ensure the tightness of π^ε when $\varepsilon \to 0$?
2. In these cases, how is made a limiting value π^0?
3. Can we choose a decreasing sequence ε_t in order to have a convergence in probability?

4 Topology Preservation

We believe that to get rigourous results, the topology preservation must not be considered as a discrete notion.

Considering two metric spaces $(\mathbb{X}, d)$ and $(\mathbb{T}, D)$, it is natural to say that an application $W : \mathbb{X} \longrightarrow \mathbb{T}$ is topology preserving if W is an homeomorphism (bicontinuous) on its image $W(\mathbb{X})$.

In the case where $\mathbb{X}$ and $\mathbb{T}$ are compact manifolds of the same dimension in Euclidean spaces $\mathcal{E}$ and $\mathcal{F}$, then we may assume that W^{-1} is defined on $\mathbb{T}$.

Let σ be a neighbourhood function defining the topology of $\mathcal{E}$ (for instance $\sigma(x, y) = \frac{1}{ad(x,y)+1}$, but there are infinitely many choices) and μ a probability on $\mathbb{T}$. We can define a continuous organized map considering a limit when the number of units n goes to ∞. Assuming that μ has a *p.d.f.* f , thanks to the cost function associated to the SOM, which here is well-defined, such a continuous organized map would minimize:

$$\mathbf{C}(W) = \int_{\mathbb{X}^2} \sigma(x, y) \|W(x) - W(y)\|_{\mathbb{T}}^2 |J_W(y)| f(W(y)) dy dx$$

where dx is the Lebesgue measure on $\mathbb{X}$, $\|.\|_{\mathbb{T}}$ is the geodesic distance on $\mathbb{T}$, W is a one to one $\mathcal{C}^1$ function and J_W its Jacobian.
So what could be a topology preserving Self Organized Map? It would only be defined in the limit $n \to \infty$ (n number of units): let $W^n(i_k), k = 1, .., n$ the map obtained with n units (according to σ). Then we interpolate $W^n(x)$ on the Voronoï tesselation defined by the $W^n(i_k)$. If it could be proved that this family of functions is uniformly continuous, then limiting values would exist. Choosing W^∞ such a limiting point of $(W^n, n \geq n_0)$, we may adress the question of topology preservation: is W^∞ "topology preserving", that means is W^∞ an homeomorphism ? Does W^∞ minimizes the cost function $\mathbf{C}$.

5 Statistical Point of View

When dealing with real data, we only observe a (large) sample of size N. If we assume for sake of simplicity that it behaves like a family of *i.i.d.* random variables, we may hope to apply the Law of Large Numbers and some asymptotic normality result.

We begin with the *L.G.N.* and the *a.s.* convergence. We assume that the data set is the realization of *i.i.d.* random variables $\{\omega_1, \cdot, \omega_N\}$, with common distribution μ. In the case of the quantization (0-neighbour) Pollard ([8]) has proved an *a.s.* convergence theorem. Let $x^*(N)$ the value of the best n-quantizer for the N-data set ($x^*(N) = argmin \sum_{i=1}^{n} \min_{1\leq j\leq p} \|\omega_i - x_j\|^2$). Let x^* be the best n-quantizer for distribution μ.
When N is large, the quantizer $x^*(N)$ is an approximation of x^*. We have the following result

Theorem 6 *Law of large numbers*
Assume the uniqueness of x^*, then $x^*(N) \longrightarrow x^* \qquad a.s.\ n \to \infty$

The asymptotic normality also holds thanks to some additional asumptions. Pollard ([9]) proved the following central limit theorem:

Theorem 7 *Central limit theorem*
Assume the uniqueness of x^* and $x^*(N)$, assume that μ has a continuous density $f(x)$ dominated by some integrable function $g(\|x\|)$, then: $\lim_{n\to\infty} n^{1/2}(x^*(N) - x^*) =_{\mathcal{L}} \mathcal{N}(0, \Sigma)$

The matrix Σ is supposed to be positive, which is satisfied here.

We may reasonably hope to prove the same results when dealing with the SOM. But there is a main difficulty: in the case of SOM, x^* does not realize the minimum of a simple cost function. So that a direct extension ot the Pollard proofs is not possible. As for most of the mathematical treatments of the Kohonen SOM, the general theorems do not apply and a specific study is needed ...

6 Conclusion

In this short paper, we have reviewed some mathematical questions that would be of interest to answer. We have selected those which were the clearest to formulate and may be the easiest to prove. It remains a lot of questions which are waiting for some light. For instance, everybody prefers to obtain the actual quantizers, but still preserve a topological organization. The way to do it is very simple: one begins the Kohonen algorithm with a strong neighbourhood function and then relaxes this function until it reaches the 0-neighbour function. What is the speed of relaxation to get a "good" SOM and quantizer has no theoretical answer at this time. Even a study of the one-dimensional SOM would give some good hints on the way to proceed in larger applications.

References

[1] Benaïm M., Fort J.C., Pagès G., Convergence of the one-dimensional Kohonen algorithm. Adv. in Appl. Probab. 30 (1998), no. 3, 850-869.

[2] Bouton C., Pagès G., Convergence in distribution of the one-dimensional Kohonen algorithms when the stimuli are not uniform. Adv. in Appl. Probab. 26 (1994), no. 1, 80–103.

[3] Cottrell M., Fort J.C., Etude d'un processus d'auto-organisation. Ann. Inst. H. Poincaré Probab. Statist. 23 (1987), no. 1, 1–20.

[4] Delattre S., Fort J.C., Pagès G., Local distortion and μ-mass of the cells of one dimensional asymptotically optimal quantizers. Comm. Statist. Theory Methods 33 (2004), no. 5, 1087-1117.

[5] Fort J.C., Pagès G., On the a.s. convergence of the Kohonen algorithm with a general neighborhood function. Ann. Appl. Probab. 5 (1995), no. 4, 1177–1216.

[6] Fort J.C., Pagès G., Asymptotics of optimal quantizers for some scalar distributions. J. Comput. Appl. Math. 146 (2002), no. 2, 253–275.

[7] Kohonen, Teuvo, Self-organization and associative memory. Springer Series in Information Sciences, 8. Springer-Verlag, Berlin, 1984. 255 pp.

[8] Pollard, David, Strong consistency of k-means clustering. Ann. Statist. 9 (1981), no. 1, 135-140.

[9] Pollard, David, A central limit theorem for k-means clustering. Ann. Probab. 10 (1982), no. 4, 919-926.

[10] Ritter H., Asymptotic level for a class of vector quantization process. IEEE Trans on Neural Networks 2 (1991), no 1, 173-175. 195-198.

[11] Ritter, H., Schulten, K., On the stationary state of Kohonen's self-organizing sensory map. Biol. Cybernet. 60 (1988), no. 1, 59-71

[12] Sadeghi A., Convergence in distribution of the multi-dimensional Kohonen algorithm. J. Appl. Probab. 38 (2001), no. 1, 136-151.