Convergence in distribution of the multidimensional Kohonen algorithm

Ali A. Sadeghi
Fachbereich Mathematik, Universität Kaiserslautern
67653 Kaiserslautern,Germany.
E-mail: sadeghi@mathematik.uni-kl.de

Abstract

Here we consider the Kohonen algorithm with a constant learning rate as a Markov process evolving in a topological space.It is shown that the process is an irreducible and aperiodic T-chain, regardless of the dimension of both data space and network and the special shape of the neighborhood function.Moreover the validity of Deoblin's condition is proved. These imply the convergence in distribution of the process to a finite invariant measure with a uniform geometric rate. In addition we show the process is positive Harris recurrent, which enables us to use statistical devices to measure its centrality and variability as the time goes to infinity.

1 Introduction

The Kohonen self-organizing map was introduced in 1982 as a simple model of a process of self-organization between different areas of the cortex and different sensory inputs [9, 10]. The simplicity of its learning algorithm immediately made the map interesting for many artificial intelligence applications. Beside the feedforward neural networks, which are designed for supervised learning, the Kohonen net is nowadays the most well known neural network with a reasonable capability for unsupervised learning of the topological features[9].

The network consists of a set of neurons, denoted here by I, which are labelled from 1 to N.To every neuron i, a set of neighboring neurons $V_i \subset I$ is assigned, such that $i \in V_i$

[0]**AMS 1991 subject classification** Primary 60J05; Secondary 60j20,92B20.
[0]**Key words and phrases.** neural networks, multidimensional Kohonen algorithm, Markov process, stochastic stability, uniform ergodicity.

and $i \in V_j$ iff $j \in V_i$. To connect the set of neurons with data space Q, which is usually a subset of $\mathbb{R}^d$, each neuron i is mapped to a weight vector $X_i \in Q$.

Every $v \in Q$ corresponds with the winner neuron $i^*(v)$ which satisfy

$$\| X_{i^*(v)} - v \| \leq \| X_i - v \| \qquad \forall \ i \in I, \tag{1}$$

where $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^d$. For the case that more than one neurons satisfy (1), we choose the neuron with lowest index as the winner.

To avoid confusions we occasionally denote the winner $i^*(v)$ corresponding to the weight vector X^t by $i^*(X^t, v)$ or i_t^*.

The adaptation of weight vectors will be carried out by an unsupervised learn process. At the end of the learn process the map $v \in Q \mapsto i^*(v) \in \{1, \cdots, N\}$ preserves (in some sense) the topology of the data space. The learning starts with an initial weight vector $X^0 = (X_1^0, \cdots, X_N^0)^T$, which will be updated as follows.

$$X_i^{t+1} = X_i^t + \epsilon_t f_{i^* i}(v^t - X_i^t) \qquad \forall \ i \in I, \quad t = 0, 1, \cdots, \tag{2}$$

where $v^t \in Q$ is an independent random variable distributed identically by some probability distribution P, $\epsilon_t \in (0, 1)$ is the so-called learning rate and $f : (i, i^*) \mapsto f_{ii^*} \in [0, 1]$ is the neighborhood function. The initial weight vector X^0 may be chosen at random from Q^N.

The functions f, in their most general form, are non-increasing with respect to neighborhood relations. To make clear what it means, let V_i^k denotes the set of neurons j for which there exist a chain of indices $(i_0, i_1, \cdots, i_k) \subset I$ such that $i_0 := i$, $i_1 \in V_i \setminus i$, $i_2 \in V_{i_1} \setminus i_1, \cdots, i_k \in V_{i_{k-1}} \setminus i_{k-1}, j \in V_{i_k} \setminus i_k$ and $j \in V_i^k$ only if $j \notin V_i^{k-1}$. Now the non-increasing function f satisfies, $f_{ik} \geq f_{il}$ if $k \in V_i^r$ and $l \in V_i^{r+1}$ for $r = 1, \cdots$.

Throughout the paper we always assume that Q is a convex subset of $\mathbb{R}^d$, $d \geq 1$. Moreover we suppose $suppP$ has a nonempty interior and $suppP \subset Q$. The distribution P is always assumed to be continuous, that is, for every set $A \subset Q$ with Lebesgue measure zero we have $P(A) = 0$.

To assess the performance of the algorithm, one needs to answer following questions. First, what is the probability of a topology preserving map takes place? Of course this question can be answered only if we have a clear definition of the topology preserving states. Second, what happens after reaching the topology preserving state? Is it possible that the topology preservation get lost? and what is the final state of the network, if the iterating algorithm converges.

To answer these questions two approaches are conceivable. One can consider the iterative equation (1) and (2) as a discrete time stochastic dynamical system on Q^N. The

performance of the algorithm is directly connected to the asymptotic behavior of this system. The major difficulty arising in this connection is that the right hand side of (2),as a function of X^t and v^t, is not continuous. The existing knowledge about such kind of systems are poor, so that drawing a practical conclusion about the Kohonen algorithm using system theoretical approaches seems to be very difficult. For the most recent works in this direction we cite [6].

The other approach to tackle the problem is to consider that $(X^t)_{t\in\mathbb{N}}$ defines a Markov process on Q^N which is homogeneous if and only if ϵ_t is time independent.

Despite the fact that in all implementations of the algorithm the data space has dimensionalities greater than one, nearly all the existing rigorous results concerning the behavior of Kohonen algorithm deal with the one-dimensional network implemented on a one-dimensional set of data. This is mainly because in the one-dimensional case the set of topology preserving states can be defined easily and it is absorbing, so that the evolution of the process can be divided into two phases, namely, *self-organization* and *convergence.*

The attempt to investigate the mathematical behavior of the algorithm in its one-dimensional setting began by T.Kohonen in [9]. M.Cottrel and J.C. Fort provided the first rigorous proof of the self-organization and convergence (in a conditional sense) in a special case of the one-dimensional algorithm[3].Since then many works have been appeared which studied different special cases of the one-dimensional algorithm in both phases, see [1, 2, 5, 7, 8] and references therein. The most resent results in this aspect are presented in [14, 15, 16]. In [15] a proof of self-organization is presented for the general one-dimensional algorithm, that is, a one-dimensional algorithm with general type of stimuli distribution, neighborhood function and learning rate, see also [16]. The convergence phase of the general one-dimensional algorithm is studied in [14], where the well known Kushner-Clark theorem is employed to show that a cooperative and irreducible differential equation governs the asymptotic behavior of the algorithm.This leads to an almost sure convergence result for the algorithm. The fact that the differential equation is cooperative, is an important difference between the one-dimensional and multidimensional cases.

Concerning the multidimensional Kohonen algorithm a well-defined performance evaluation and a rigorous mathematical study are obviously lacking. The later is the subject of the present work.

In this paper we consider the Kohonen algorithm as a Markov process on a topological space X. We take to account neither the dimensionalities of the network or data space nor the special shape of neighborhood function or the distribution of data in Q. We do not answer to the questions stated above directly. Instead of that we study the weak convergence of the algorithm with a constant learning rate. The one-dimensional Kohonen algorithm with constant learning rate were already studied in [2].

Our aim is to show that the process *settles down*, or converges, to a stable or stationary regime independent of its initial starting point. A necessary condition for the occurrence

of such a asymptotical behavior is the existence of a finite invariant measure. The role of finite invariant measure is to describe the final regime of the chain.The next step is to study the way that the convergence takes place.In this connection we show that the convergence occur with a geometric rate. And finally it is worth to know more about the characters of the invariant measure. This will be done by using the central limit theorem, the law of large numbers and the law of iterated logarithm, through which we measure centrality and variability of X^t as t becomes large.

Methodology. The principal tools which we apply come from the theory of stochastic stability of Markov processes, see [11]. It turns out that we deal with a aperiodic T-chain and this implies that the process is irreducible and positive Harris recurrent.These, together with the validity of Deoblin's condition provide the necessary basis for the ergodic properties of our interest.

2 Preliminaries

In this section we review the basic definitions and results from the theory of discrete time homogeneous Markov processes which will be used in the Section 3. A detailed treatment of these ideas can be found in [11].

Let X be a topological space and $(X^t)_{t\in\mathbb{N}}$ a $(X, \mathcal{B}(X), \mathcal{P})-$valued stochastic process. $(X^t)_{t\in\mathbb{N}}$ is said to be a *Markov process* if

$$\mathcal{P}(X^{k+1} \in A | X^k) = \mathcal{P}(X^{k+1} \in A | X^t, t = 0, \cdots, k),$$

for all $k \in \mathbb{N}$ and $A \in \mathcal{B}(X)$.

We adopt the following notation in this paper

$$\begin{aligned} P^t(x, A) &:= \mathcal{P}(X^t \in A \quad | \quad X^0 = x), \\ P(x, A) &:= P^1(x, A), \\ L(x, A) &:= \mathcal{P}(\cup_{t=1}^{\infty}[X^t \in A] \quad | \quad X^0 = x), \\ Q(x, A) &:= \mathcal{P}(\cap_{t=1}^{\infty} \cup_{k=t}^{\infty} [X^k \in A] \quad | \quad X^0 = x). \end{aligned}$$

The following condition will be referred as *Doeblin's condition.*

Condition : There is a probability measure φ on $\mathcal{B}(X)$, an integer $m \geq 1$ and $\epsilon < 1, \delta > 0$ such that

$$\inf_{x\in X} P^m(x, A) > \delta \quad \text{if} \quad \varphi(A) > \epsilon.$$

This version of Doeblin's condition differs slightly from the one stated in classical books like[4], however it is easy to show that every process which is Doeblin in the classical sense is also Doeblin in this sense.

The Markov process $(X^t)_{t\in\mathbb{N}}$ is called *φ-irreducible* if there exists a measure φ on $\mathcal{B}(X)$ such that, whenever $\varphi(A) > 0$, we have $L(x, A) > 0$ for all $x \in X$.

If $(X^t)_{t\in\mathbb{N}}$ is φ-irreducible, then there exists an integer d and a collection of disjoint sets $D^1, \cdots, D^d \subset X$ (a "d-cycle") with the property that

$$P(x, D^{i+1}) = 1 \quad \text{for all } x \in D^i \quad P(x, D^1) = 1 \quad \text{for all } x \in D_d,$$

and the set $[\cup_{i=1}^d D^i]^c$, the complement of $\cup_{i=1}^d D^i$, is φ-null. The chain is *aperiodic* if the largest d for which a d-cycle occurs is equal to one.

A kernel T will be called a *continuous component* of a function $K : (X, \mathcal{B}(X)) \to \mathbb{R}_+$ if

(*i*) For $A \in \mathcal{B}(X)$ the function $T(., A)$ is lower semi-continuous.

(*ii*) For each $x \in X$, $T(x, .)$ is a substochastic measure on $\mathcal{B}(X)$ and $K(x, A) \geq T(x, A)$ for all $A \in \mathcal{B}(X)$.

The continuous component T is called *non-trivial* at x if $T(x, X) > 0$.

Let a be a probability measure on $\mathbb{Z}_+$ and define the Markov transition function K_a as

$$K_a := \sum_{t=1}^{\infty} a(t) P^t.$$

If $(X^t)_{t\in\mathbb{N}}$ is a Markov process such that,for some a, K_a admits a continuous component T which is non-trivial for all $x \in X$, then $(X^t)_{t\in\mathbb{N}}$ is said to be a $T-chain$.

A $\sigma-$finite measure π on $\mathcal{B}(X)$ is called *invariant* if

$$\pi(A) = \int P(x, A)\pi(dx) \quad \text{for all} \quad A \in \mathcal{B}(X).$$

If $(X^t)_{t\in\mathbb{N}}$ is φ-irreducible and $Q(x, A) = 1$ whenever $\varphi(A) > 0$, then $(X^t)_{t\in\mathbb{N}}$ is called *Harris recurrent*. It is shown in [13] that for a Harris recurrent chain there exists an essentialy unique invariant measure π. If X^t is Harris recurrent and π is finite , then $(X^t)_{t\in\mathbb{N}}$ is called *positive Harris recurrent*.

The process $(X^t)_{t\in\mathbb{N}}$ will be called *bounded in probability* if for each $x \in X$ and each $\epsilon > 0$, there exists a compact set $K \subset X$ such that

$$\liminf_{t\to\infty} P^t(x, K) \geq 1 - \epsilon.$$

Lemma 1 is proved in [12].

Lemma 1. *Suppose that* $(X^t)_{t\in\mathbb{N}}$ *is an irreducible* $T-$*chain. Then* $(X^t)_{t\in\mathbb{N}}$ *is positive Harris recurrent if and only if* $(X^t)_{t\in\mathbb{N}}$ *is bounded in probability.*

A set $A \in \mathcal{B}(X)$ is ν_m-small if there exist an $m > 0$ and a non-trivial measure ν_m on $\mathcal{B}(X)$ such that for all $x \in A, \quad B \in \mathcal{B}(X)$,

$$P^m(x, B) \geq \nu_m(B).$$

$A \in \mathcal{B}(X)$ is ν_a- petite if there exists a non-trivial measure ν_a on $\mathcal{B}(X)$ such that

$$K_a(x, B) \geq \nu_a(B),$$

for all $x \in A, \quad B \in \mathcal{B}(X)$.

Lemma 2. *Suppose* $(X^t)_{t\in\mathbb{N}}$ *is an aperiodic* $\varphi-$*irreducible T-chain. Then every compact subset of* X *is small.*

Proof. This is a direct consequence of Theorem 6.0.1 and Theorem 5.5.7. of[11]. □

In this paper we consider convergence of the Kohonen algorithm in term of its transition probabilities and also along its sample paths. For this we endow the set of measures on $\mathcal{E}$ with the total variation norm defined as

$$\|\mu\| := \sup |\int f\mu(dx)|,$$

where supremum is taken over all measurable functions $f : X \to \mathbb{R}$, such that $|f(x)| \leq 1$ for all $x \in X$. It can be shown that

$$\|\mu\| = \sup_{A\in\mathcal{E}} \mu(A) - \inf_{A\in\mathcal{E}} \mu(A).$$

The convergence of measures on the topological space X in total variation norm also implies their weak convergence, that is , if $\|\mu_k - \mu\| \to 0$, then for every $f \in C(X)$ we have

$$\lim_{k\to\infty} \int f d\mu_k = \int f d\mu.$$

A chain $(X^t)_{t\in\mathbb{N}}$ is called *uniformly ergodic* if there exists an invariant probability measure π such that

$$\sup_{x\in X} \|P^t(x, .) - \pi(.)\| \to 0, \quad t \to \infty.$$

It is known that a necessary and sufficient condition for an aperiodic Markov chain $(X^t)_{t\in\mathbb{N}}$ to be uniformly ergodic is that it satisfies the Doeblin's condition, see [11].

The class of uniformly ergodic chains has many interesting statistical properties, including the validity of the centeral limit theorem and the law of large numbers.

3 Basic properties of the Kohonen algorithm

In this section we establish some basic results concerning the properties of Kohonen's algorithm, where $(X^t)_{t\in\mathbb{N}}$, as defined by (1) and (2), is considered as a discrete time Markov processes evolving in the topological space $X = Q^N$. $Q \subset \mathbb{R}^d$ is a convex set which contains $\bar{suppP}$ (the closure of $suppP$). In case that $suppP$ is convex, it is easy to show that if $X^0 \in [suppP]^N$, then $X^t \in [suppP]^N$, for all t and one can set $Q = \bar{suppP}$. However if P has a non-convex support, the event $X^t \notin [suppP]^N$ may happen for some t and this situation may even remain unchanged for a long time. In this case Q must be a convex set such that $suppP \subset Q$.

Before going to technicalities, let us remind that the chain is not Feller. This is because in cases that $x \in Q^N$ has some coinciding components,the stimuli in any arbitrarily small neighborhood of x can be assigned to different winner neurons, that is, $P(x,.)$ is not lower semi-continuous if x has coinciding components .

The following conventions will be used in this section. $D_1 := \{x \in Q^N | x_i \neq x_j,\ i \neq j\}$ is the set of weight vectors with pairwise distinct components.

$$D_2 := \{x \in Q^N \quad | \quad x_i \neq x_j \quad \text{for some} \quad i \quad \text{and all} \quad j \quad \neq i\}$$

is the set of weight vectors with at least one distinct component, $D'_1 := Q^N \setminus D_1$ and $D'_2 := Q^N \setminus D_2$.

$$B(x,\eta) := \{y \in Q^N \quad | \quad \|x_i - y_i\| \leq \eta \quad \forall\, i\}$$

is N pair of closed balls with radius η which is centered at a. $[a]$ denotes the integer part of a real number a. The Voronoi tessellation of Q induced by $x \in Q^N$, is the family $\{C^i(x), \quad i \in I\}$, where $C^i(x)$ is defined as follows.

$$C^i(x) = \{v \in Q \quad | \quad i^*(x,v) = x_i\}.$$

In this paper we consider $(D_2, \mathcal{B}(D_2), \mathcal{P})$ as our probability space. The reason for this choice is that the Kohonen algorithm consists a T-chain and has reachable points on D_2, see Lemma 6 and Proposition 1. However, compared to $(Q^N, \mathcal{B}(Q^N), \mathcal{P})$ this probability space has the disadvantage that D_2 is not compact.

3.1 Continuous component

Let $\mathcal{V}_n(x,A)$ be the set of all events $\nu = (v^0, \cdots, v^{n-1})$ which take $X^0 = x \in Q^N$ to $A \in \mathcal{B}(Q^N)$. Note that $P^n(x,A) = P^{\otimes n}(\mathcal{V}_n(x,A))$. We remind that P is always assumed to be continuous.

Lemma 3. *For any* $x \in D_1$ *and almost all* $\nu \in \mathcal{V}_n(x, A)$ *there exists a* $\eta > 0$ *such that* $\nu \in \mathcal{V}_n(y, A)$ *for all* $y \in B(x, \eta)$.

Proof. If A includes no interior point, then $\mathcal{V}_n(x, A)$ has Lebesgue measure zero and the lemma is trivial.

Next suppose A includes some interior points. In this case without loss of generality we assume that A is open.

Let $X^1, \cdots, X^n$ be the weight vectors corresponding to $\nu \in \mathcal{V}_n(x, A)$ and

$$j(X^t, v^t) := \arg \min_{I \setminus i^*(X^t, v^t)} \|X_i^t - v^t\|$$

the second nearest neuron to v^t. If there exists more than one $j(X^t, v^t)$, any of them may be chosen.

For $x \in D_1$ there exists a $\eta > 0$ such that

$$\eta \leq 0.5(\|X^t_{j(X^t, v^t)} - v^t\| - \|X^t_{i^*(X^t, v^t)} - v^t\|) \quad \forall \quad t = 0, \cdots, n-1. \tag{3}$$

Note that the set of the events ν, for which there dose not exists such a $\eta > 0$ has Lebesgue measure zero.

We show, by induction, that if $Y^0 = y \in B(x, \eta)$, then after the event ν we have $Y^n \in B(X^n, d^{0.5n}\eta)$.

Let $Y^{t-1} \in B(X^{t-1}, \eta)$ for some $t \in \{1, \cdots, n\}$. We have

$$Y_i^t = (1 - \epsilon_t f_{ii^*}) Y_i^{t-1} + \epsilon_t f_{ii^*} v^{t-1} \quad \forall \quad i \in I. \tag{4}$$

(3) implies $i^*(v^{t-1}, Y^{t-1}) = i^*(v^{t-1}, X^{t-1})$ and therefore

$$X_i^t - (1 - \epsilon_t f_{ii^*})\eta < Y_i^t < X_i^t + (1 - \epsilon_t f_{ii^*})\eta \quad \forall \quad i \in I,$$

that is, $Y^t \in B(X^t, d^{0.5}\eta)$ where d is the dimension of the data space.

Recall that A is an open subset of Q^N, for sufficiently small η if $Y^0 \in B(X^0, \eta)$, then after the event ν we have $Y^n \in B(X^n, d^{0.5n}\eta) \subset A$. □

Let $\mathcal{V}_n^\alpha(x, A) \subset \mathcal{V}_n(x, A)$ be the set of all events $(v^0, \cdots, v^{n-1})$ taking all $y \in B(x, \alpha)$ to A. We have $\mathcal{V}_n^\beta(x, A) \subset \mathcal{V}_n^\alpha(x, A)$ if $\beta > \alpha$. Furthermore from the proof of the Lemma 3 it follows that for any $\epsilon > 0$ there exists an $\alpha > 0$ such that $P^{\otimes n}(\mathcal{V}_n(x, A) \setminus \mathcal{V}_n^\alpha(x, A)) < \epsilon$.

Lemma 4. *For all* $t \in \mathbb{N}$, $P^t(x, A)$ *is a continuous function of* x *on* D_1.

Proof. Consider a point $x \in D_1$ and any arbitrary sequence $y_n \in B(x, \alpha_n)$ such that $y_n \to x$, $\alpha_n \to 0$ as $n \to \infty$. Without loss of generality we assume $y_n \in D_1$ for all n.

We need to show that $P^t(y_n, A) \to P^t(x, A)$ as $n \to \infty$, or equivalently $P^{\otimes t}(\mathcal{V}_t(y_n, A) \setminus \mathcal{V}_t(x, A)) \to 0$ and $P^{\otimes t}(\mathcal{V}_t(x, A) \setminus \mathcal{V}_t(y_n, A)) \to 0$. $y_n \in B(x, \alpha_n)$ and $x \in B(y_n, \alpha_n)$ implies that $\mathcal{V}_t^{\alpha_n}(x, A) \subset \mathcal{V}_t(y_n, A)$ and $\mathcal{V}_t^{\alpha_n}(y_n, A) \subset \mathcal{V}_t(x, A)$, respectively.Now $\mathcal{V}_t(x, A) \setminus \mathcal{V}_t(y_n, A) \subset \mathcal{V}_t(x, A) \setminus \mathcal{V}_t^{\alpha_n}(x, A)$, and $\mathcal{V}_t(y_n, A) \setminus \mathcal{V}_t(x, A) \subset \mathcal{V}_t(y_n, A) \setminus \mathcal{V}_t^{\alpha_n}(y_n, A)$ which together with $P^{\oplus t}(\mathcal{V}_t(x, A) \setminus \mathcal{V}_t^{\alpha_n}(x, A)) \to 0$ and $P^{\oplus t}(\mathcal{V}_t(y_n, A) \setminus \mathcal{V}_t^{\alpha_n}(y_n, A)) \to 0$ as $n \to \infty$ leads us to the required result.

□

Remark 1. It can be shown that the Lemma is not valid for $x \in D_1'$.

For $x \in Q^N$, $A \in \mathcal{B}(Q^N)$ we next introduce

$$\begin{aligned} \mathcal{H}(x) &=: \{v \in Q \mid x_{i^*(x,v)} \neq x_j \quad \forall \quad j \in I \setminus i^*(x, v)\}, \\ \bar{P}(x, A) &=: \mathcal{P}\{X^1 \in A \mid X^0 = x \quad \text{and} \quad v \in \mathcal{H}(x)\}. \end{aligned}$$

$\bar{P}(x, A)$ is the probability of $X^1 \in A$ if $X^0 = x$ and the winner neuron has a distinct weight vector. This function has the following properties:

- For $x \in D_1$ we have $\bar{P}(x, A) = P(x, A)$,
- For $x \in D_2$ $\bar{P}(x, A)$ defines a substochastic measure on $\mathcal{B}(D_2)$, $\bar{P}(x, A) \leq P(x, A)$ and $\bar{P}(x, D_2) > 0$.
- For $x \in D'_2$ we have $\bar{P}(x, Q^N) = 0$.

Sequelly we will show that the function $\bar{P}(x, A)$ defines a continuous component for some K_a.

Lemma 5. *For any* $A \in \mathcal{B}(Q^N)$ *the function* $\bar{P}(x, A)$ *is lower semi-continuous on* D_2.

Proof. Let $x \in D_2$ consider any sequence $x^k \to x$, $x^k \in D_2$. We need to show that $\underline{\lim}_{k\to\infty} \bar{P}(x^k, A) \geq \bar{P}(x, A)$.

First note that the Lebesgue measure of $\mathcal{H}(x) \setminus \mathcal{H}(x^k)$ tends to zero as $k \to \infty$.

Introduce

$$\begin{aligned}\Gamma^k &:= \{v \in \mathcal{H}(x^k) \quad | \quad X^0 = x^k, X^1 \in A\},\\ \Gamma_1^k &:= \{v \in \mathcal{H}(x) \quad | \quad X^0 = x^k, X^1 \in A\},\\ \Gamma_2^k &:= \{v \in \mathcal{H}(x^k) \setminus \mathcal{H}(x) \quad | \quad X^0 = x^k, X^1 \in A\}.\end{aligned}$$

We have $\bar{P}(x^k, A) = P(\Gamma_1^k) + P(\Gamma_2^k)$.

If $x_i \neq x_j$ for all $j \in I \setminus i$, then it is easy to see that the Lebesgue measure of

$$(C^i(x^k) \setminus C^i(x)) \cup (C^i(x) \setminus C^i(x^k))$$

tends to zero, as $k \to \infty$. This ensures,

$$\lim_{k\to\infty} P(\Gamma_1^k) = P(\{v \in \mathcal{H}(x) \quad | \quad X^0 = x, X^1 \in A\}) = \bar{P}(x, A).$$

Now, the fact that $P(\Gamma_2^k) \geq 0$ for all k implies $\underline{\lim}_{k\to\infty} \bar{P}(x^k, A) \geq \bar{P}(x, A)$.

□

$\bar{P}(x, A)$ is a lower semi-continuous function and for any fixed $x \in Q^N$. it defines a measure on D_2. So it can be considered as a substochastic transition kernel on $(D_2, \mathcal{B}(D_2))$ and its $n-$step transition probability kernel may be defined in the classical manner, see [11], such that the function $\sum_{t=0}^{\infty} \bar{P}^t(x, A)a(t)$ which we are going to use in the Lemma 6, is well defined.

Lemma 6. *The process* $(X^t)_{t\in\mathbb{N}}$ *is a T-chain.*

Proof. Consider the function

$$T(x, A) := \sum_{t=0}^{\infty} \bar{P}^t(x, A)a(t), \quad x \in D_2, A \in \mathcal{B}(D_2), \tag{5}$$

where $a(t) := (1 - \theta)\theta^t$ and $0 < \theta < 1$. It is clear that $K_a(x, A) \geq T(x, A)$, and moreover $T(x, D_2) > 0$ for all $x \in D_2$. The lower semi-continuity of $T(x, A)$ follows form Lemma 5.

□

Remark 2. In the definition of $T(x, A)$ one could use any other probability measure on $\mathbb{Z}^+$ instead of $a(t) := (1 - \theta)\theta^t$.

3.2 Irreducibility and Positive Harris recurrence

From now on we always assume that ϵ_t is a constant.

While the results established so far are valid without any restriction on the dimensionalites of the network and the data space, the results presented in the rest of this section are valid only if $Q \subset \mathbb{R}^d,\ d > 1$.

Proposition 1 shows that for a reasonably large class of neighborhood functions there exists infinite number of reachable points in D_2. Here Ω_{x^*} denotes a neighborhood of the point x^*.

Proposition 1. *(i) Let $y^* \in suppP$ with $y_i^* = y_j^*$ for all i, j. Then $L(x, \Omega_{y^*}) > 0$ for all $x \in Q^N$.*
(ii)Consider a point $x^* \in suppP$ wich has exactly $N-1$ coinsiding components, that is, $x_k^* \neq x_i^*$ for some k and all $i \neq k$, $x_i^* = x_j^*$ for all $i, j \neq k$. Suppose there exists a non-empty set of neurons $\tau \subset I$ such that $f_{kj} = 0$ for all $j \in \tau$ and for any $i \in I \setminus k$ there exists a $j \in \tau$ such that $f_{ij} \neq 0$. Then $L(x, \Omega_{x^*}) > 0$ for all $x \in Q^N$.

Note that $y^* \notin D_2$, but $x^* \in D_2$.

Proof. *(i)* For any t the event $v^t \in \Omega_{y^*},\ t = 1, \cdots, n$ happens with a positive probability. After each step some of the neurons lay closer to Ω_{y^*} while the others remain fixed.So if all the different neurons become winner in some times (which happens with positive probability), then for sufficiently big n we have $X^n \in \Omega_{y^*}$.

(ii) Choose a neighborhood $\Omega_{x_k^*}$ and a set $\Omega \subset Q$ such that $\Omega_{x_k^*} \times \Omega^{N-1} \subset \Omega_{x^*}$.

Using *(i)* we can assume that $x \in [\Omega_{x_k^*}]^N$. Moreover without los of generality we assume $x \in [\Omega_{x_k^*}]^N \cap D_1$, such that every neuron may win with a positive probability.(Note that ever $x \in D'_1$ may be taken to D_1 within a finite number of steps with probability one.)

Consider the event $v_t \in \Omega_{x_k^*}$ and $i^*(X^t, v_t) = k,\ t = 1, \cdots, n$. For sufficiently large n there exists a ball $B(a, r) \subset \Omega_{x_k^*}$ such that $x_j \in B(a, r)$ for all $j \in \Psi_k$ and $x_j \notin B(a, r)$ for all $j \in I \setminus \Psi_k$. Remind that $\tau \subset I \setminus \Psi_k$, there exist a chain of events, $i^*(X^t, v_t) \in \tau, t = n+1, \cdots, l$, such that the winner get always closer to Ω without changing the position of k.

This means after the event $i^*(X^t, v_t) \in \tau,\ t = n+1, \cdots, m$ for sufficiently large n we have $x_k \in \Omega_{x_k^*},\ x_j \subset \Omega$ for all $j \neq k$.

□

In fact, Proposition 1 is also valids for decreasing learning rates provided $\sum_{t=0}^{\infty} \epsilon_t = \infty$.

Next we introduce a measure $T(.)$ on $\mathcal{B}(D_2)$, such that all set of $\mathcal{B}(D_2)$ with a positive $T(.)-$measure can be reached by X^t, no matter where is the starting point.

Lemma 7. *$(X^t)_{t\in\mathbb{N}}$ is a $T(x^*,.)-$irreducible process where x^* is any fixed point defined as in Proposition 1(ii).*

Proof. Let $T(x^*,A)>0$. The lower semi-continuity of $T(.,A)$ implies there exists a neighborhood O_{x^*} such that $T(y,A)>0$ for all $y\in O_{x^*}$. Now for any $x\in D_2,\quad A\in\mathcal{B}(D_2)$ we have

$$\begin{aligned}\sum_{t=0}^{\infty}P^t(x,A)>K_a(x,A) &\geq \int_{O_{x^*}}K_a(x,dy)K_a(y,A)\\ &\geq \int_{O_{x^*}}K_a(x,dy)T(y,A)>0,\end{aligned}$$

which ensures $L(x,A)>0$.

□

Lemma 8. *If Q is bounded, then the process $(X^t)_{t\in\mathbb{N}}$ is positive Harris recurrent.*

Proof. This is a direct implication of Lemma 1,6,7. It is clear from definition of the algorithm that $(X^t)_{t\in\mathbb{N}}$ is bounded in probability.

□

3.3 Uniform ergodicity

In order to establish the desired ergodicity result, we need to study the aperiodicity of the process and also the validity of Deoblin's condition. This will be end in this section.

Apriodicity

First we need to establish the Proposition 2.

Proposition 2. *The folowing statements hold.*
(i) For all $x\in Q^N$ and $v^0\in Q$ we have $P(v \mid i^*(X^0,v)=i^*(X^1,v))>0$.
(ii) Let $X^0=x$ and $Y^0=y$. For each t if $P(v \mid i^*(X^t,v)=i^*(Y^t,v))>0$ then $P(v \mid i^*(X^{t+1},v)=i^*(Y^{t+1},v)$ and $i^*(X^t,v)=i^*(Y^t,v))>0$.

Proof. Observe that after each adaptation step, with v^t and X^t, the winner neuron is again the neuron with minimum distance to the stimulus of the step in question, that is, $i^*(X^t, v^t) = i^*(X^{t+1}, v^t)$. This is the case as long as $f_{ij} \leq f_{ii}$ for all i and j.

It turns out that for almost all $v^t \in Q$ there exists a neighborhood Ω_{v^t} such that if $v \in \Omega_{v^t}$, then $i^*(X^t, v) = i^*(X^{t+1}, v)$, which proves both (i) and (ii).

□

Lemma 9. *The process* $(X^t)_{t\in\mathbb{N}}$ *is aperiodic.*

Proof. It will be enough if we show that for any $x \in Q^N$ there exists a Borel set $A \subset D^i$ and $n \in \mathbb{N}$ such that $P^n(x, A) > 0$ and $P^{n+1}(x, A) > 0$.

Let $r \in \{1, \cdots, N\}$, $l := [i/d]$. We denote by $A(r) = [a_{ij}(r)]_{Nd\times Nd}$ and $B(r) = [b_{ij}(r)]_{Nd\times Nd}$ the $Nd \times Nd$ matrices with $a_{ij}(r) = 0, b_{ij}(r) = 0$ for $i \neq j$, and $a_{ij}(r) = 1 - \epsilon f_{lr}, b_{ij}(r) = \epsilon f_{lr}$ for $i = j$. $V(t)$ is a $Nd \times 1$ vector with $V_i(t) := v_i^t$ for $1 \leq i \leq d$ and $V_{i+d}(t) := V_i(t)$ for $i + d \leq Nd$.

Now the algorithm (1) and (2) may be rewritten as

$$\begin{aligned} X^t = {} & A(i^*_{t-1}) \cdots A(i^*_0)X^0 + A(i^*_{t-1}) \cdots A(i^*_1)B(i^*_0)V(0) \\ & + \cdots + A(i^*_{t-1})B(i^*_{t-2})V(t-2) + B(i^*_{t-1})V(t-1), \end{aligned} \tag{6}$$

where i^*_t is the winner at time t.

Next let $X^0 = x$. After an arbitrary event ν_0 we define $y := X^1$ and Y^t is the corresponding chain with $Y^0 = y$. Now consider the set of events

$$\begin{aligned} \Gamma = \{(v^0, v^1, \cdots, v^{n-1}) \quad | \quad & i^*(x, v^0) = i^*(y, v^0),\ i^*(X^1, v^1) = i^*(Y^1, v^1) \\ & , \cdots, i^*(X^{n-1}, v^{n-1}) = i^*(Y^{n-1}, v^{n-1})\}. \end{aligned}$$

The Proposition (2) implies Γ happens with a positive probability.

Using (6) after the event Γ we have

$$X^t - Y^t = A(i^*_{t-1}) \cdots A(i^*_0)(X^0 - Y^0).$$

All eigenvalues of the matrices $A(r)$, $r = 1, \cdots, N$ are positive and smaller than $1 - \epsilon$. This implies for any given η there exists t such that $Y^t \in B(X^t, \eta)$.

We come to the conclusion that for sufficiently big t, if D^i posses an open subset A and $X^t \in A$, then we have $Y^t \in A$. In other words, there exists $A \subset D^i$ and $n \in \mathbb{N}$ such that $P^n(x, A) > 0$ and $P^{n+1}(x, A) > 0$.

□

Deoblin's Condition

In this section we show that the Doeblin's condition is valid on D_2. This result can be easily generalized to whole Q^N. The validity of Doeblin's condition strengthens the $\varphi-$irreducibility of the algorithm. From applications point of view, it amount to the existence of a minimum positive probability for reaching any given set after some finite number of times, regardless of the initial situation.

Set

$$D_2^\gamma := \{x \in Q^N | \quad \|x_i - x_j\| \geq \gamma \quad \text{for some} \quad i \quad \text{and all} \quad j \neq i\}$$

For any $\gamma > 0, D_2^\gamma$ is a compact subset of D_2.

Proposition 3. *Suppose* $f_{ii} > f_{ij}$ *for all* $j \neq i$. *There exists* $\gamma > 0, \quad \delta_1 > 0$ *such that for all* $x \in D_2$,

$$P(x, D_2^\gamma) > \delta_1.$$

Proof. There exist a neuron i and a constant $\theta > 0$ such that $P(i^* = i) \geq 1/N$ and $\sup_{a \in Q} P(B(a, \theta)) < 1/N$. Now consider the set $\mathcal{E} := \{v \in Q \quad | \quad i^*(x, v) = i \quad \text{and} \quad \|v - x_i\| \geq \theta\}$. The event $v_0 \in \mathcal{E}$ happens with the probability greater than $\delta_1 := 1/N - \theta > 0$. It is enough to show that if $\gamma \leq \epsilon(f_{ii} - f_{ij})\theta$ and $v_0 \in \mathcal{E}$, then $X^1 \in D_2^\gamma$.

$v_0 \in \mathcal{E}$ implies $\|v_0 - X_j^0\| \geq \|v_0 - X_i^0\|$ and

$$X_j^1 = X_j^0 + \epsilon f_{ij}(v_0 - X_j^0),$$

for all j. An easy geometrical consideration yields

$$\|X_j^1 - X_i^1\| \geq \epsilon(f_{ii} - f_{ij})\|v_0 - X_i^0\| \geq \epsilon(f_{ii} - f_{ij})\theta \geq \gamma \quad \forall \quad j \neq i,$$

which ensures $X^1 \in D_2^\gamma$.

□

Lemma 10. *The process* $(X^t)_{t \in \mathbb{N}}$ *satisfies Deoblin's condition.*

Proof. First we consider a set D_2^γ, for a $\gamma > 0$ such that Proposition 3 is valid. D_2^γ is compact, so Lemma 2 together with Lemmas 6,7,9 can be used to conclude that it is a small set, that is, there exists an $m > 0$, and a non-trivial measure v_m on D_2, such that for all $x \in D_2^\gamma, \quad B \in \mathcal{B}(D_2)$,

$$P^m(x, B) \geq v_m(B).$$

Now Proposition 3 implies for all $x \in D_2, \quad B \in \mathcal{B}(D_2)$,

$$P^{m+1}(x, B) \geq \delta_1 v_m(B).$$

□

4 Main results

In this section we state the main results of the paper, namely Theorem 1 and Theorem 2. Theorem 1 establishes the uniform ergodicity of the chain and some different stability properties which are equivalent to the uniform ergodicity.This result will be completed by sample paths results in Theorem 2.

Theorem 1. *The following equivalent statements hold.*

(i) The process $(X^t)_{t\in\mathbb{N}}$ *is uniformly ergodic on* D_2.

(ii)There exists $0 < r < 1$ *and* $R < \infty$ *such that for all* $x \in D_2$.

$$\|P^t(x,.) - \pi(.)\| \leq Rr^t.$$

(iii) D_2 *is* v_m*-small for some* m *and in particular*

$$\|P^t(x,.) - \pi(.)\| < (1 - v_m(D_2))^{t/m}, \quad t \to \infty.$$

Proof. This is a direct result of Theorem 16.0.2 of [11] and Lemmmas 9,10. □

Remark 4. By the definition of the total variation norm we have

$$|P^t(x,.) - \pi(.)| \leq \|P^t(x,.) - \pi(.)\|$$

and therefore the results of Theorem 1 remain valid if we replace $\|.\|$ by $|.|$ in *(i),(ii)* and *(iii)*.

As mentioned in Section 2, the convergence in total variation norm in a topological space implies the weak convergence of the measures in question. Thus the asymptotic properties established in Theorem 1 are much stronger than those one could expect from a Feller chain, even if it possesses a unique invariant measure.

The result established in (iii) can be interpreted as follows. Although the way that the process behaves depends on the initial point x, but there exists a minimal level of independency from x, which will be enjoyed by the $m-$skeleton chain. This minimal level of independency is measured by ν_m.

Knowing that the process converges in distribution to a probability measure $\pi(.)$, the natural question which now arises is what are the characteristics of $\pi(.)$. The Theorem 2 provides measures of centrality and variability of X^t for large t, which strengthen the convergence result of Theorem 1.

Theorem 2. *(i) The low of large numbers holds for any function* $g : D_2 \mapsto \mathbb{R}$ *satisfying* $\pi(|g|) := \int |g(x)|\pi(dx) < \infty$, *that is,*

$$\lim_{t\to\infty} \frac{1}{t} S_t(g) = \pi(g)$$

where

$$S_t(g) := \sum_{k=1}^{t} g(X^k).$$

(ii) Suppose $g^2(x) \le 1$ *for all* $x \in D_2$, *and let* $\bar{g} := g - \pi(g)$. *Then the constant*

$$\gamma_g^2 := E_\pi[\bar{g}^2(X^0)] + 2\sum_{k=1}^{\infty} E_\pi[\bar{g}(X^0)\bar{g}(X^k)]$$

is well defined, non-negative and finite , and

$$\lim_{t\to\infty} \frac{1}{t} E_\pi[(S_t(\bar{g}))^2] = \gamma_g^2.$$

(iii) Under the same conditions as in (ii), if $\gamma_g^2 = 0$ *then*

$$\lim_{t\to\infty} \frac{1}{t^{1/2}} (S_t(g))^2 = 0$$

(iv) Under the same conditions as in (ii), if $\gamma_g^2 > 0$ *then the central limit theorem and law of iterated logarithm holds, that is,*

$$\lim_{t\to\infty} \mathbf{P}_x\{(t\gamma_g^2)^{-1/2} S_t(\bar{g}) \le a\} = \int_{-\infty}^{a} \frac{1}{(2\pi)^{1/2}} e^{-x^2/2} dx$$

and for each $X^0 = x \in D_2$ *the limit infimum and limit supremum of the sequence*

$$(2\gamma_g^2 t \log\log(t))^{-1/2} S_t(\bar{g})$$

are respectively -1 and +1 with probability one.

Proof. This is a direct consequense of Lemma 8 and Theorem 17.0.1 of [11]. □

The variance given in (iv) measures the magnitude of variability of X^t as $t \to \infty$.

5 Conclusion

This work presents a systematic study of those features of the Kohonen algorithm which are not a result of either special dimensionalities of data space and the network or the special shape of the neighborhood function. In fact to date no special implementation of the algorithm is reported, in which either of the mentioned aspects play a key roll. The facts that we used can be summarized as follows.

1- The very basic fact, which we take to consideration is that in each step of the learning phase the only information needed are those of the last step; all the older information can be forgotten. This is a typical situation in which a Markov process analysis comes to considerations.

2- To use the results of the theory of homogeneous Markov chains, we assume a constant learning rate. This may differ from the actual implementations of the algorithm, but by no way is it a strong restriction. The results presented here show that a constant step Kohonen algorithm can be used to achieve a degree of self-organization. To improve the produced map, a smaller learning rate then can be used in further learning steps.This is in fact a strategy which can help to improve the performance of the algorithm in many implementations.

3- One of the features of the algorithm which made its analysis complicated, is the discontinuity of $P(x,.)$. However this function is almost every where continuous. This fact is used in this paper for defining the function $\bar{P}(x,.)$, which is actually a non-trivial lower semi-continuous lower bound for $P(x,.)$.

4- The kohonen algorithm, regardless of its dimensionalities, always have infinite number of reachable point. Any point of Q^N, which has $N-1$ coinciding components is a reachable point, in the sense that any neighborhood of it can be reached from any arbitrary initial state with a positive probability.However the set of points with $N-1$ coinciding components have a Lebesgue measure zero.

Acknowledgement. The Author would like to thank Professor D. Prätzel-Wolters for his support wich made this work possible.

References

[1] Bouton, C., & Pagés, G. (1993) Self-organization and a.s. convergence of the one-dimensional Kohonen algorithm with non-uniformly distributed stimuli.*Stochastic Procces and their Applications* , **47** , 249-274 .

[2] Bouton, C., & Pagés, G. (1994) Convergence in distribution of the one-dimensional Kohonen algorithm when the stimuli are not uniform . *Advances in Applied Probability*, **26**, 80-103 .

[3] Cottrell,M., & Fort,J.C. (1987). Étude d'un processus d'auto-organisation . *Annales de l' Institut Henri Poincaré*,**23**(1) , 1-20.

[4] Doob,J.L., (1953) *Stochastic Processes.* John Wiley & Sons, New York.

[5] Erwin, E., Obermeyer,K. & Schulten ,K.(1992). Self-organizing maps: ordering, convergence properties and energy functions.*Biological Cybernetic* ,**67** , 47-55 .

[6] Fang, H.,Gong,G.& Qian,M. (1997). Annealing of iterative stochastic schemes, *SIAM J. Control Optm,* 35, pp. 1886-1907.

[7] Flanagan,J.A. (1996). Self-organisation in Kohonen's SOM,*Neural Networks* ,**7** , 1185-1197.

[8] Fort, J. C. and Pagés, G. (1995). *On the a.s. convergence of the Kohonen algorithm with a general neighborhood function*, The Ann. of Appl. Prob. , 4 , pp. 1177-1216.

[9] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps.*Biological Cybernetics*,**43**, 59-69.

[10] Kohonen,T. (1995). *Self-organizing maps* , Springer.

[11] Meyn, S.P., Tweedie, R.L. (1993).*Markov chains and stochastic stability*, Springer.

[12] Meyn, S.P., Tweedie, R.L. (1992). Stability of Markovian processes 1: Criteria for discrete-time chains.*Adv. Appl. Prob.*,**24**,542-574.

[13] Orey, S. (1971).*Limit theorems for Markov chain transition probabilities.* Van Nostrand Reinhold Mathematical Studies, London.

[14] Sadeghi, A.A.(1998). Asymptotic behavior of self-organizing maps with non-uniform stimuli distribution, The Ann. of Appl. Prob., **8**,281-299.

[15] Sadeghi A.A., Self-Organization property of Kohonen's map with general type of stimuli distribution , preprint 181, Arbeitsgruppe Technomathematik, Universität Kaiserslautern, September 1997, Accepted for publication in Neural Networks.

[16] Sadeghi A.A.,(1998) Self-organization and convergence of the one-dimensional Kohonen algorithm, Proc. of the ESANN98 conference,173-178, Brussels, D Facto ed. .