Communicated by Klaus Obermayer

# On the Distribution and Convergence of Feature Space in Self-Organizing Maps

**Hujun Yin**
**Nigel M. Allinson**
*Image Engineering Laboratory, Department of Electronics, University of York, York YO1 5DD, United Kingdom*

**In this paper an analysis of the statistical and the convergence properties of Kohonen's self-organizing map of any dimension is presented. Every feature in the map is considered as a sum of a number of random variables. We extend the Central Limit Theorem to a particular case, which is then applied to prove that the feature space during learning tends to multiple gaussian distributed stochastic processes, which will eventually converge in the mean-square sense to the probabilistic centers of input subsets to form a quantization mapping with a minimum mean squared distortion either globally or locally. The diminishing effect, as training progresses, of the initial states on the value of the feature map is also shown.**

## 1 Introduction

The self-organizing map (SOM) (Kohonen 1984) has been studied for over a decade. There are still many aspects to be exploited. Even the most general theory concerning this algorithm is far from complete or is lacking in vigorous mathematical explanation, as Kohonen (1991) and other researchers have remarked (Erwin *et al.* 1992a,b; Bauer and Pawelzik 1992). The ordering and the convergence of the SOM have been proved by Kohonen (1984) and Cottrell and Fort (1986) for a one-dimensional chain of neurons with a one-step neighborhood function. Erwin *et al.* (1992a,b) extended the proof of the Kohonen chain's ordering and convergence from the one-step neighborhood function to any monotonically decreasing function centered on the winning neuron. However, for high-dimensional maps and dimensional reduction problems, the convergence and ordering are very difficult to examine or even to describe. By considering the SOM's Markovian properties, Ritter and Schulten (1988) derived a Fokker–Planck equation to describe the convergence of the feature space in the vicinity of equilibrium.

Here, the learning dynamics of the algorithm are studied using probability and statistics theories as we consider each neuron's weight, or

 

feature, as a stochastic process, which consists of a sum of random variables with time-varying scalars. Each feature consists of two parts, contributions from the initial states and from the input data. As the training progresses, the contribution from the initial states to the *value* of the feature map is shown to tend to zero. The second contribution is proved through an extended Central Limit Theorem to tend to a gaussian process, and to converge in the mean-square (m.s.) sense to the probabilistic density center of an input subset. The neighborhood relationship is a function of time, winning neurons and their spatial relationship to the other neurons. Its implicit relationship with the winning neurons often makes explicit analysis of the learning process difficult. This problem, however, has been overcome in the following analysis.

## 2 The SOM Algorithm and Its Rewritten Form

The algorithm uses a set of neurons, $\mathbf{Y}$ of $M$-dimension, to form a topology conserving (partially or globally) discrete mapping of an $N$-dimensional input space, $\mathbf{X} \in \mathbf{R}^N$. Every neuron, indexed by $c \in \mathbf{Y}$, is connected, in parallel, to all dimensional components of input sample, $\mathbf{x} \in \mathbf{X}$, $\mathbf{x} = [x_1, x_2, \ldots x_N]^T$. The connection strengths, or weights, are $\mathbf{w}_c(n) = [w_{c1}(n), w_{c2}(n), \ldots w_{cN}(n)]^T$, $c \in \mathbf{Y}$, where $n$ is the discrete time step, and $n \geq 0$. The initial weights are randomly set. During training, at each time step, $n$, a randomly selected input sample, $\mathbf{x}(n)$, from the input space $\mathbf{X}$ is presented to the network. Every neuron compares its weights with this input; and the best-matching neuron, or winner, indexed by $\nu(n)$, can be found using

$$\nu(n) = \arg\min_{c \in \mathbf{Y}} \{\| \mathbf{x}(n) - \mathbf{w}_c(n) \|\} \tag{2.1}$$

The weights are then updated according to the following rule

$$\mathbf{w}_c(n+1) = \mathbf{w}_c(n) + \alpha(n) h(c, \nu, n) [\mathbf{x}(n) - \mathbf{w}_c(n)] \qquad \forall c \in \mathbf{Y} \tag{2.2}$$

or, since in most cases only *scalar-valued* $\{\alpha(n)\}$ and $\{h(c, \nu, n)\}$ terms are used, this can be expressed as

$$\begin{aligned} w_{ci}(n+1) &= w_{ci}(n) + \alpha(n) h(c, \nu, n) [x_i(n) - w_{ci}(n)], \\ & \qquad i = 1, 2, \ldots N; \qquad \forall c \in \mathbf{Y} \end{aligned} \tag{2.3}$$

where $h(c, \nu, n)$ is termed the neighborhood function and depends on time $n$, neuron $c$, and winner $\nu$. There are many types of such functions. A stepped function is used in our analysis (other forms of neighborhood functions will also be mentioned), namely

$$h(c, \nu, n) = \begin{cases} 1, & \text{if } c \in \mathbf{N}_\nu(n) \\ 0, & \text{if } c \notin \mathbf{N}_\nu(n) \end{cases} \tag{2.4}$$

where $\mathbf{N}_\nu(n)$ is the neighborhood set centered on the winner. $\mathbf{N}_\nu(n)$ *"should be very wide in the beginning (of the training) and shrink monotonically with time"* until the winner is the only member of the neighborhood set. *"A good global ordering"* may then be formed (Kohonen 1990).

The adaptation gain coefficients $\{\alpha(n), n \geq 0\}$ are *scalar-valued, decrease monotonically,* and satisfy the following conditions (Kohonen 1984):

$$0 < \alpha(n) < 1, \qquad \lim_{n\to\infty} \sum \alpha(n) \to \infty, \quad \lim_{n\to\infty} \sum \alpha^2(n) < \infty \tag{2.5}$$

The third condition of 2.5 has been replaced by Ritter and Schulten (1988) with a less restrictive one, namely, $\lim_{n\to\infty} \alpha(n) \to 0$.

Equation 2.3 can be rewritten as

$$\begin{aligned} w_{ci}(n+1) &= w_{ci}(0) \prod_{k=0}^{n} [1 - \alpha(k)h(c,\nu,k)] \\ &\quad + \sum_{k=0}^{n} x_i(k) \prod_{l=k+1, k<n}^{n} [1 - \alpha(l)h(c,\nu,l)]\, \alpha(k)h(c,\nu,k), \\ &\quad i = 1,2,\ldots N; \qquad \forall c \in \mathbf{Y} \end{aligned} \tag{2.6}$$

## 3 The Effect of Initial States

To examine the first term of 2.6, i.e., the contribution of the initial states, we write

$$b_{ci}(n) \equiv \prod_{k=0}^{n} [1 - \alpha(k)h(c,\nu,k)] = \prod_{k=0, c\in\mathbf{N}_\nu(k)}^{n} [1 - \alpha(k)] \tag{3.1}$$

Only if the neuron, $c$, is in the neighborhood set, $\mathbf{N}_\nu(n)$, at time $n$, will its weights be modified and the corresponding terms will appear in 3.1. Let $D_c(m) \equiv \{$*the number of time steps for which c is not in* $\mathbf{N}_\nu(n)$ *beginning at time* $m$, $m = 0,1,\ldots n; n \geq 0\}$ represent the intervals between updates of the neuron $c$'s weights. For each neuron $c$, there exists an integer $\lambda_c$, for which $D_c(m) < \lambda_c$, $m = 0,1,\ldots\infty$. $\lambda_c$ will be a finite number, otherwise the neuron $c$ will not fire again (we need to assume that there are no "dead" neurons). Hence

$$b_{ci}(n) = \prod_{m=0, m=m+D_c(m)}^{n} [1 - \alpha(m)] \tag{3.2}$$

Taking natural logarithms of both sides gives

$$\begin{aligned} \ln b_{ci}(n) &= \sum_{m=0, m=m+D_c(m)}^{n} \ln [1 - \alpha(m)] \\ &\leq \sum_{m=0, m=m+D_c(m)}^{n} [-\alpha(m)] \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\lambda_c} \sum_{m=0, m=m+D_c(m)}^{n} \lambda_c \alpha(m) \\
&= -\frac{1}{\lambda_c} \sum_{m=0, m=m+D_c(m)}^{n} \overbrace{[\alpha(m) + \alpha(m) + \cdots + \alpha(m)]}^{\lambda_c} \\
&\leq -\frac{1}{\lambda_c} \sum_{k=0}^{n} \alpha(k) \qquad (3.3)
\end{aligned}
$$

The last inequality holds because $\{\alpha(k)\}$ decreases monotonically. From the second condition of 2.5, we obtain

$$
b_{ci}(n) \leq \exp\left[-\frac{1}{\lambda_c} \sum_{k=0}^{n} \alpha(k)\right] \xrightarrow{n \to \infty} 0 \qquad (3.4)
$$

Thus the first term of equation 2.6, i.e., the contribution from the initial states, will tend to zero if the initial states are finite. The second term of 2.6 will not depend on the initial states although $h(c, \nu, n)$ is a function of the winning neurons of each iteration including the first. As we will show in the next section, this second contribution will converge to the centroids of the subsets of the input data, which do not contain the randomly selected initial states.

Equation 3.4 shows that the *values* of the final features will not be affected by the initial states *provided the adaptation gains satisfy the conditions* 2.5. We use *values* here to avoid confusion with the *order* of the map, as in some circumstances the initial states may affect the order, or both order and value, of the final states due to inappropriate implementation of the neighborhood function and/or the adaptation gains.

The above results apply only for stepped neighborhood functions. However, they can be easily extended for general convex neighborhood functions. For any such functions, it holds that $\inf\{h(c, \nu, n)\} \leq h(c, \nu, n) \leq \sup\{h(c, \nu, n)\}$, hence

$$
\begin{aligned}
b_{ci}(n) &\equiv \prod_{k=0}^{n} [1 - \alpha(k) h(c, \nu, k)] \\
&\leq \prod_{k=0, c \in \mathbf{N}_\nu(k)}^{n} [1 - \inf\{h(c, \nu, k)\} \alpha(k)] \qquad (3.5)
\end{aligned}
$$

Taking logarithms of both sides gives a similar result to equation 3.3 except for a factor of $\inf\{h(c, \nu, k)\}$, which will not affect the result: $b_{ci}(n) \to 0$.

## 4 The Probabilistic Distribution of the SOM Feature Space

As the effect of initial states will tend to zero, the feature map will depend primarily on the second term of 2.6, i.e., the contribution from the input

space. Since the input vectors are drawn randomly, or independently, from the input set $\mathbf{X}$, then the second contribution can be treated as a weighted sum of independent random variables (r.v.s) $\{\mathbf{x}(n),\ n \geq 0\}$. Each neuron receives inputs from a set, termed $\mathbf{X}_c(n)$, which is a time-varying subset of the input set $\mathbf{X}$. At the beginning of the training phase, all subsets are maximally overlapped with each other. As the training progresses and the neighborhood size shrinks to just one neuron, input subsets $\{\mathbf{X}_c(n), c \in \mathbf{Y}, n \geq 0\}$ will eventually be mutually separated with

$$\bigcup_{c\in\mathbf{Y}} \mathbf{X}_c(n) \xrightarrow{n\to\infty} \mathbf{X}, \qquad \text{and} \qquad \mathbf{X}_c(n) \bigcap \mathbf{X}_{c'}(n) \xrightarrow{n\to\infty} \phi,$$
$$c \neq c',\ \forall c,\ c' \in \mathbf{Y} \tag{4.1}$$

As time tends to infinity, $\{\mathbf{X}_c(n)\}$ will tend to $\{\mathbf{X}_c\}$, which are termed the final input subsets.

Suppose the probability density function of the input set $\mathbf{X}$ is $f(\mathbf{x})$, the probability of an input sample, $\mathbf{x}(n)$, belonging to a subset, $\mathbf{X}_c(n)$, is given by

$$P(\mathbf{X}_c, n) = \int_{\mathbf{x}\in\mathbf{X}_c(n)} f(\mathbf{x})\, d\mathbf{x} \qquad \forall c \in \mathbf{Y} \tag{4.2}$$

and within each input subset $\mathbf{X}_c(n)$, the probability density function is

$$f_c(\mathbf{x}, n) = f(\mathbf{x})/P(\mathbf{X}_c, n) \qquad \forall c \in \mathbf{Y} \tag{4.3}$$

As time tends to infinity, $\{\mathbf{X}_c(n), P(\mathbf{X}_c, n), \text{and } f_c(\mathbf{x}, n), c \in \mathbf{Y}\}$ will tend to $\{\mathbf{X}_c, P(\mathbf{X}_c), \text{and } f_c(\mathbf{x}), c \in \mathbf{Y}\}$, respectively. So $f_c(\mathbf{x}) = f(\mathbf{x})/P(\mathbf{X}_c)$.

The Central Limit Theorem is concerned with the statistical properties of a sum of independent r.v.s. The differences in the present case are that such a sum (cf. the second term of 2.6) is a sum of *time-varying* weighted r.v.s, where the variance of each weighted random variable will tend to zero, rather than a finite number. In the following, we will show that the variance of the sum of these weighted r.v.s will also tend to zero (otherwise the algorithm will not converge). So we cannot directly apply any existing version of the Central Limit Theorem to this analysis. It is necessary to extend the theorem to this particular application. We introduce an extended form of the theorem, the proof of which is given in the Appendix.

**Theorem 4.1.** *If $\{X_n, n \geq 0\}$ are independent r.v.s with finite means of $\{m_n, n \geq 0\}$, finite variances of $\{\sigma_n^2, n \geq 0\}$, and finite higher moments, i.e., for any $\delta > 0$,*

$$\mu_n^{(2+\delta)} = \int_{X_n} X_n^{2+\delta} f(X_n)\, dX_n \ < \infty \tag{4.4}$$

*where $f(X_n)$ is the density function of $X_n$. $\{a_k(n), k = 0, 1, \ldots n,\ n \geq 0\}$ is a set of time-varying real numbers, which satisfy*

$$\text{(i) } 0 < a_k(n) < 1; \quad \text{(ii) } \sum_{k=0}^{n} a_k(n) \xrightarrow{n\to\infty} 1; \quad \text{(iii) } \sum_{k=0}^{n} a_k^2(n) \xrightarrow{n\to\infty} 0 \tag{4.5}$$

*The weighted sum* $\{\sum_{k=0}^{n} a_k(n)X_n,\ n \geq 0\}$ *will tend to a gaussian distributed process with means of* $\{m(n) = \sum_{k=0}^{n} a_k(n)m_k,\ n \geq 0\}$ *and variances of* $\{\sigma^2(n) = \sum_{k=0}^{n} a_k^2(n)\sigma_k^2,\ n \geq 0\}$, *and with* $m(n) \to E\{m_n\}$, $\sigma^2(n) \to 0$ *when* $n \to \infty$. *Furthermore, if* $X_n \to X'$, *then such a weighted sum will converge in the m.s. sense to m, the mean of* $X'$.

The second term of 2.6 is a time-varying weighted sum of independent r.v.s. and the time-varying weight set $\{a_k(n), k = 0, 1, \ldots n;\ n \geq 0\}$ is given by

$$a_k(n) \equiv \left\{ \prod_{l=k+1, k<n}^{n} [1 - \alpha(l)h(c, \nu, l)] \right\} \alpha(k)h(c, \nu, k) \tag{4.6}$$

Next we shall prove that this set will satisfy the three conditions of 4.5. The first condition, $0 < a_k(n) < 1$, holds because of 2.4 and 2.5; and the second one holds because

$$\begin{aligned}
\sum_{k=0}^{n} a_k(n) &= \sum_{k=0}^{n} \left\{ \prod_{l=k+1, k<n}^{n} [1 - \alpha(l)h(c, \nu, l)] \right\} \alpha(k)h(c, \nu, k) \\
&= \{1 - [1 - \alpha(n)h(c, \nu, n)]\} \\
&\quad + [1 - \alpha(n)h(c, \nu, n)] \{1 - [1 - \alpha(n-1)h(c, \nu, n-1)]\} \\
&\quad + \cdots \\
&\quad + [1 - \alpha(n)h(c, \nu, n)] [1 - \alpha(n-1)h(c, \nu, n-1)] \\
&\quad \cdots [1 - \alpha(1)h(c, \nu, 1)] \{1 - [1 - \alpha(0)h(c, \nu, 0)]\} \\
&= 1 - \prod_{k=0}^{n} [1 - \alpha(k)h(c, \nu, k)]
\end{aligned} \tag{4.7}$$

From Section 3, the second term of the above equation will tend to zero.

For the last condition, considering

$$\sum_{k=0}^{n} a_k^2(n) \equiv \sum_{k=0}^{n} \left\{ \prod_{l=k+1, k<n}^{n} [1 - \alpha(l)h(c, \nu, l)]^2 \right\} \alpha^2(k)h^2(c, \nu, k) \tag{4.8}$$

Since $\sum_{k=0}^{\infty} \alpha^2(k)$ converges, so for any arbitrary small value $\varepsilon$, there exists a number $\kappa$, for which $\sum_{\kappa}^{\infty} \alpha^2(k) < \varepsilon$, and because $0 < [1 - \alpha(l)h(c, \nu, l)] < 1$, then

$$\begin{aligned}
\lim_{n \to \infty} \sum_{k=\kappa}^{n} a_k^2(n) &= \sum_{k=\kappa}^{\infty} \left\{ \prod_{l=k+1}^{\infty} [1 - \alpha(l)h(c, \nu, l)]^2 \right\} \alpha^2(k)h^2(c, \nu, l) \\
&< \sum_{k=\kappa}^{\infty} \alpha^2(k) < \varepsilon
\end{aligned} \tag{4.9}$$

For a finite $\kappa$, since $\sum_{k=0}^{\infty} \alpha(k)$ diverges, $\sum_{k=\kappa}^{\infty} \alpha(k)$ will also diverge, and from Section 3, $\prod_{l=\kappa+1}^{\infty} [1 - \alpha(l)h(c, \nu, l)]$ will also tend to zero. Since

$\sum_{k=0}^{\kappa} \alpha^2(k) < \theta$, (a constant), therefore

$$\begin{aligned}\lim_{n\to\infty} \sum_{k=0}^{\kappa} a_k^2(n) &= \sum_{k=0}^{\kappa} \left\{ \prod_{l=k+1}^{\infty} [1 - \alpha(l)h(c,\nu,l)]^2 \right\} \alpha^2(k) h^2(c,\nu,k) \\ &< \left\{ \prod_{l=\kappa+1}^{\infty} [1 - \alpha(l)h(c,\nu,l)]^2 \right\} \sum_{k=0}^{\kappa} \alpha^2(k) \\ &< \theta \prod_{l=\kappa+1}^{\infty} [1 - \alpha(l)h(c,\nu,l)] \to 0 \end{aligned} \tag{4.10}$$

We conclude from 4.9 and 4.10 that the last condition also holds. The above results, together with the nearest neighbor matching law of the algorithm, result in a lemma:

**Lemma 4.1.** The feature space of the SOM algorithm is approximate gaussian distributed stochastic processes, and will converge in the m.s. sense to the centroids of the final input subsets,

$$\mathbf{w}_c(n) \overset{n\to\infty}{\longrightarrow} \mathbf{m}_c = \frac{1}{P(\mathbf{X}_c)} \int_{\mathbf{X}_c} \mathbf{x} f(\mathbf{x})\, d\mathbf{x}, \qquad \forall c \in \mathbf{Y} \tag{4.11}$$

where $\{\mathbf{m}_c\}$ is termed the final feature space, and is the set of cluster centers of the final input subsets $\{\mathbf{X}_c\}$. Each final subset $\mathbf{X}_c$ has hyperplane boundaries which are defined by

$$\| \mathbf{x} - \mathbf{m}_c \| = \| \mathbf{x} - \mathbf{m}_{c'} \| \qquad \forall c' \in \mathbf{Y}, \text{ but } c' \neq c \tag{4.12}$$

## 5 Conclusions

We have analyzed the statistical properties of the feature space of the SOM algorithm. From the proof of its gaussian distribution approximation we have also formally proved the convergence of the SOM algorithm. An extension to the relaxed conditions on adaptive gains has also been proved (Yin and Allinson 1993). The Lemma 4.1 means that the SOM will eventually satisfy two necessary conditions for minimizing the mean squared distortion of vector quantization. The results are dimension independent, i.e., the convergence exists for any dimensionality. Provided that the shrinking speeds of the neuron neighborhood are adequate then, at least, local topological ordered maps will be formed.

## Appendix: Proof of Theorem 4.1

First consider the zero-mean case, i.e. $\{m_n = 0, n \geq 0\}$.

The following two formulas can easily be obtained

$$e^{jX} = 1 + jX - \frac{X^2}{2} + \beta\frac{|X|^{2+\delta}}{2^\delta}, \qquad \forall X \in \mathbf{R}, \qquad |\beta| \le 1 \tag{A.1}$$

$$1 - X = e^{-X} - \frac{\beta' X^2}{2}, \qquad \forall X \ge 0, \quad 0 < \beta' < 1 \tag{A.2}$$

In A.1 let $X = a_k(n)X_k\omega$ and taking the expectation of both sides, the characteristic function of $a_k(n)X_k$ is obtained

$$\begin{aligned}\Phi_k(\omega, n) &= E\{e^{j\omega a_k(n)X_k}\} \\ &= 1 - \frac{a_k^2(n)\sigma_k^2}{2}\omega^2 + \beta_k\frac{a_k^{2+\delta}(n)\mu_k^{(2+\delta)}}{2^\delta}|\omega|^{2+\delta}, \qquad |\beta_k| \le 1\end{aligned} \tag{A.3}$$

Let $X = a_k^2(n)\sigma_k^2\omega^2/2$ in A.2, then

$$1 - \frac{a_k^2(n)\sigma_k^2}{2}\omega^2 = e^{-\{[a_k^2(n)\sigma_k^2]/2\}\omega^2} - \frac{\beta_k'}{2}\left(\frac{a_k^2(n)\sigma_k^2\omega^2}{2}\right)^2, \qquad 0 < \beta_k' < 1 \tag{A.4}$$

Thus we can write

$$\Phi_k(\omega, n) = e^{-\{[a_k^2(n)\sigma_k^2]/2\}\omega^2}(1 + \gamma_k) \tag{A.5}$$

where

$$\gamma_k = e^{\{[a_k^2(n)\sigma_k^2]/2\}\omega^2}\left[\beta_k\frac{a_k^{2+\delta}(n)\mu_k^{2+\delta}|\omega|^{2+\delta}}{2^\delta} - \frac{\beta_k'}{2}\left(\frac{a_k^2(n)\sigma_k^2\omega^2}{2}\right)^2\right]$$

Since $a_k(n) \overset{n\to\infty}{\longrightarrow} 0$, thus $a_k^2(n)\sigma_k^2\omega^2 < 1$ holds for any finite area of $\omega$, and since

$$\begin{aligned}\left(\frac{a_k^2(n)\sigma_k^2\omega^2}{2}\right)^2 &= \left(\frac{a_k^2(n)\sigma_k^2\omega^2}{2}\right)^{1-\delta/2}\left(\frac{a_k^2(n)\sigma_k^2\omega^2}{2}\right)^{1+\delta/2} \\ &< \frac{1}{2^{1-\delta/2}}\left(\frac{a_k^2(n)\sigma_k^2\omega^2}{2}\right)^{1+\delta/2} \\ &< \frac{1}{4}a_k^{2+\delta}(n)\mu_k^{(2+\delta)}|\omega|^{2+\delta}\end{aligned} \tag{A.6}$$

The last inequality holds because $(\sigma_k^2)^{2+\delta} \le (\mu_k^{(2+\delta)})^2$. So the following inequality holds

$$|\gamma_k| < \frac{9}{8}e^{1/2}a_k^{2+\delta}(n)\mu_k^{(2+\delta)}|\omega|^{2+\delta} \tag{A.7}$$

Since $\{X_n, n \ge 0\}$ are independent r.v.s, then

$$\begin{aligned}\Phi(\omega, n) &= \mathrm{E}\left\{e^{j\omega\sum_{k=0}^n a_k^2(n)X_k}\right\} = \prod_{k=0}^{n}\Phi_k(\omega, n) \\ &= e^{-(\omega^2/2)\sum_{k=0}^n a_k^2(n)}(1 + \gamma_1)(1 + \gamma_2)\cdots(1 + \gamma_n)\end{aligned} \tag{A.8}$$

The error of its gaussian distribution approximation is

$$\begin{aligned}\left|\Phi(\omega,n)-e^{-(\omega^2/2)\sum_{k=0}^{n}a_k^2(n)\sigma_k^2}\right| &< (1+|\gamma_1|)\,(1+|\gamma_2|)\cdots(1+|\gamma_n|)-1\\ &< e^{|\gamma_1|+|\gamma_2|+\cdots|\gamma_n|}-1\\ &= e^{\sum_{k=0}^{n}|\gamma_k|}-1\\ &< e^{2|\omega|^{2+\delta}\mu_{\max}^{(2+\delta)}\sum_{k=0}^{n}a_k^{2+\delta}(n)}-1 \xrightarrow{n\to\infty} 0 \qquad \text{(A.9)}\end{aligned}$$

because if $\sum_{k=0}^{n}a_k^2(n)\xrightarrow{n\to\infty}0$, and $\delta>0$, then $\sum_{k=0}^{n}a_k^{2+\delta}(n)\xrightarrow{n\to\infty}0$, and where $\mu_{\max}^{(2+\delta)}=\max\{\mu_n^{(2+\delta)},\ n\geq 0\}$.

From A.9 and using a lemma of Uspensky (1937) (i.e., if the characteristic function of random variable $S$ tends to the characteristic function of a gaussian distributed variable, then the distribution of $S$ tends to that gaussian function), we can conclude that $\{\sum_{k=0}^{n}a_k(n)X_k,\ n\geq 0\}$ tends to a gaussian distributed process with zero mean and variance

$$\sigma^2(n)=\sum_{k=0}^{n}a_k^2(n)\sigma_k^2<\sigma_{\max}^2\sum_{k=0}^{n}a_k^2(n)\xrightarrow{n\to\infty}0 \qquad \text{(A.10)}$$

In the nonzero mean case $\{m_n\neq 0,\ n\geq 0\}$, if every $m_n$ is a finite number, then the biased r.v.s $\{X_n',\ n\geq 0\}$ can be divided into $\{X_n+m_n,\ n\geq 0\}$, where $\{X_n\}$ are zero mean r.v.s and according to a Corollary of Slutsky's theorem (Chow and Teicher 1978) (i.e., if $\{\rho,\pi,\rho_n,\pi_n,\ n\geq 0\}$ are finite constants with $\rho_n\xrightarrow{n\to\infty}\rho$, $\pi_n\xrightarrow{n\to\infty}\pi$, and $X_n\xrightarrow{n\to\infty}X$, then $\rho_nX_n+\pi_n\to\rho X+\pi$), the weighted sum $\sum_{k=0}^{n}a_k(n)X_k'$ is also gaussian distributed with finite means $m(n)=\sum_{k=0}^{n}a_k(n)m_k(<m_{\max})$ and finite variances $\sigma^2(n)$, which will tend to zero, when $n$ tends to infinity. Furthermore, if $X_n'\xrightarrow{n\to\infty}X'$ (with the mean of $m$), then

$$m(n)=\sum_{k=0}^{n}a_k(n)m_k\xrightarrow{n\to\infty}m \qquad \text{(A.11)}$$

That is $\sum_{k=0}^{n}a_k(n)X_n'$ will converge in the m.s. sense to the mean of the $X'$. □

## Acknowledgments

The authors are grateful to the reviewers for many helpful comments.

## References

Bauer, H.-U., and Pawelzik, K. R. 1992. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Trans. Neural Networks* 3(4), 570–579.

Chow, Y. S., and Teicher, H. 1978. *Probability Theory: Independence, Interchangeability and Martinggales*. Springer-Verlag, London.

Cottrell, M., and Fort, J. C. 1986. A stochastic model of retinotopy: A self-organizing process. *Biol. Cybernet.* **53**, 405–411.

Erwin, E., Obermayer, K., and Schulten, K. 1992a. Self-organizing maps: Ordering, convergence properties and energy functions. *Biol. Cybernet.* **67**, 47–55.

Erwin, E., Obermayer, K., and Schulten, K. 1992b. Self-organizing maps: Stationary states, metastability and convergence rate. *Biol. Cybernet.* **67**, 35–45.

Kohonen, T. 1984. *Self-Organization and Associative Memory.* Springer-Verlag, London.

Kohonen, T. 1990. The Self-Organizing Map. *Proc. IEEE* **78**(9), 1464–1480.

Kohonen, T. 1991. Self-organizing maps: Optimization approaches. In *Artificial Neural Networks*, T. Kohonen, *et al.*, eds., pp. 981–990. Elsevier, Amsterdam.

Ritter, H., and Schulten, K. 1986. On the stationary states of Kohonen's self-organizing sensory mapping. *Biol. Cybernet.* **54**, 99–106.

Ritter, H., and Schulten, K. 1988. Convergence properties of Kohonen's topology conserving maps: Fluctuations, stability, and dimension selection. *Biol. Cybernet.* **60**, 59–71.

Uspensky, J. V. 1937. *Introduction to Mathematical Probability.* McGraw-Hill, New York.

Yin, H., and Allinson, N. M. 1993. *Statistical Analysis and Treatment of Kohonen's Self-Organising Map.* Tech. Rep., Image Engineering Lab., University of York, UK.

Received March 14, 1994; accepted January 20, 1995.