Pergamon

Computers Math. Applic. Vol. 33, No. 3, pp. 45–63, 1997
Copyright©1997 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0898-1221/97 $17.00 + 0.00

PII: S0898-1221(96)00236-2

Convergence Theorems for the Kohonen Feature Mapping Algorithms with VLRPs

J. F. Feng
Laboratory of Biomathematics, The Babraham Institute
Cambridge CB2 4AT, United Kingdom

B. Tirozzi
Mathematical Department, University of Rome "La Sapienza"
P.le A. Moro, 00185 Rome, Italy

(Received July 1996; accepted August 1996)

Abstract—The convergence of the Kohonen feature mapping algorithm with vanishing learning rate parameters (VLRPs) is considered, which includes the simple competitive learning algorithm as a special case. A few examples show that the learning fails to converge to "global minima," in general. Then, we present a novel approach which enables us to find out a new family of VLRPs such that the corresponding learning algorithm converges to the set of "global minima" with probability one. The new VLRPs is a generalization of the well-known rate parameters used in the simulated annealing. A numerical example is also included to confirm our theoretical approach. We believe that this discovery is of importance for a large class of learning algorithms in neural networks and statistics.

Keywords—Kohonen feature mapping algorithm, Supermartingale, Global minima, Stochastic differential equation, Vanishing learning rate parameters (VLRPs).

1. INTRODUCTION

In recent years, there are extensive research works devoted to the study of the Kohonen feature mapping algorithm, both theoretically and numerically [1–3]. In [4,5], and references given therein, the authors consider the equilibrium states of the Kohonen feature mapping algorithm with the learning rate parameter independent of time. In [4], a thorough investigation of the existence and the number of the metastable states is carried out. In [6–8], the asymptotic property of the one-dimensional Kohonen feature mapping algorithm is studied. Recently, a novel approach [9] to the problem of constructing topology preserving maps is introduced, which is based upon a Hebbian adaption rule with winner-take-all like competition. Here, we first consider the convergence problem of the Kohonen feature mapping algorithm (see [3, p. 232]) with the *nonincreasing* vanishing learning rate parameters (VLRPs) $\eta(t) > 0$, satisfying the usual restrictions found in stochastic approximation theory [10–15]

$$\int_0^\infty \eta(u)\,du = \infty, \tag{I}$$

$$\int_0^\infty \eta^2(u)\,du < \infty. \tag{II}$$

Partially supported by the CNR of Italy, J. F. was also partially supported by the A. von Humboldt Foundation of Germany.

Typeset by $\mathcal{A}_\mathcal{M}\mathcal{S}$-TEX

The constraints (I) and (II) above are usually imposed for the stochastic approximation algorithm (see, for example [10,12–14]), and the reason for a family of learning rate parameters satisfying them is fully explained in [10,14]. See also, Section 3 of the present paper, where we assert that the condition (II) is not a necessary one. Examples in Section 3 show that, in general, there are metastable states for the algorithm. Note that a Canonical candidate of $\eta(t)$ under the restrictions (I) and (II) will be

$$\eta(t) = \frac{1}{t^\alpha},$$

for $1/2 < \alpha \leq 1$ (see [3, p. 223; 15, p. 259]).

The above conclusions naturally suggest to us to ask the question: does it exist a general rule (VLRPs) for the learning algorithm which allows the system to get out of the metastable states? In other words, we look for a family of VLRPs which has a role like the decreasing 'temperature' in simulated annealing. Nevertheless, an example in Section 3 of the present paper indicates that under the constraints (I) and (II), the learning algorithm will stay at some local minima with a positive probability.

It was first noted in [15, p. 259] that in a linear learning algorithm with VLRPs the restriction (II) above is unnecessary and it could be replaced by a much weaker condition

$$\lim_{t\to\infty} \eta(t) = 0. \tag{II$'$}$$

Based upon the self-similarity property of Brownian motion and results of simulated annealing in [16], we present a novel and rigorous approach to determine a new family of VLRPs. This new family of VLRPs which is between $1/\log t$ and $1/\sqrt{\log t}$, ensures that the learning algorithm with the VLRPs escapes from the local minima and reaches the desired global minima with probability one. This fact is shown in Section 3. Note that this family of VLRPs fulfills the restriction (I) and violates the restriction (II), but it satisfies (I) and (II$'$). We believe that our discovery is of general guidance for a class of learning algorithms with VLRPs, such as the learning algorithm of Oja's law [3], Hebb learning [17], the em and EM algorithms [18], and some most recently proposed algorithms like [19,20], etc.

2. A CONVERGENCE THEOREM

2.1. Notation and Results

For a concrete description of our result, we first briefly review the Kohonen feature mapping algorithm in detail.

In the Kohonen feature mapping algorithm, there is a single layer of output units $O_i(n) \in \{1, 0\}$, $i = 1, \dots, N$ at time n, each fully connected to a set of inputs $\xi_j(n)$, $j = 1, \dots, M$, via connections $w_{ij}(n)$. In the sequel, we assume that the inputs $\xi_j(n)$, $j = 1, \dots, M$ are chosen independently according to a probability distribution P. For each presentation of the input $\xi_j(n)$, $j = 1, \dots, M$ we choose one of the output units, called the winner. The winner is the output unit with the smallest distance between its connections and the inputs

$$\|w_i(n) - \xi(n)\|,$$

for vectors $w_i(n) = (w_{ij}(n), j = 1, \dots, M)$, $\xi(n) = (\xi_j(n), j = 1, \dots, M)$, where $\|\cdot\|$ represents the Euclidean norm. Let $\bar{I}(\cdot,\cdot)$ be the function:

$$\bar{I}\left(w_i(n), \xi(n+1)\right) = I_{\{\|w_i(n)-\xi(n+1)\| < \|w_j(n)-\xi(n+1)\|, j\neq i\}}\left(w_i(n), \xi(n+1)\right),$$

where I_A is the indicator function, i.e., $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$.

The Kohonen feature mapping algorithm ensures the weights update decreasingly according to its distance with respect to the winner

$$w_{ij}(n+1) = w_{ij}(n) + \eta(n) \sum_k \Lambda(i,k) \bar{I}\left(w_k(n), \xi(n+1)\right) \cdot \left(\xi_j(n+1) - w_{ij}(n)\right), \tag{1}$$

$i = 1, \dots, N$, $j = 1, \dots, M$ or in vector form

$$w_i(n+1) = w_i(n) + \eta(n) \sum_k \Lambda(i,k) \bar{I}\left(w_k(n), \xi(n+1)\right) \cdot \left(\xi(n+1) - w_i(n)\right), \tag{2}$$

where $\eta(n)$ is the positive learning parameter $\eta(0) < 1$, $\eta(n) \geq \eta(n+1)$, and $\Lambda(i,j)$ is a nonincreasing function of $\|i-j\|$. If $\Lambda(i,j) = \delta_{ij}$, the above algorithm is called the simple competitive learning.

After the learning procedure is finished, any set of input vectors will be partitioned into nonoverlapping clusters. This means that a new incoming signal $\xi(n+1)$ is classified as the pattern i if it is closest to the weight w_i. In other words, the new signal $\xi(n+1)$ is recognized to be of the type w_i, if and only if

$$\|w_i - \xi(n+1)\| \leq \|w_j - \xi(n+1)\|, \qquad j \neq i.$$

Note that the nonlinearity of the dynamics above is addressed by the function $\bar{I}$. In the case considered in [15, p. 279], $N = 1$, and so the dynamics defined by (1) is linear because there is no competition at all. Furthermore, when $\Lambda(i,i) = 1$ and $\Lambda(i,j) = 0$ for $i \neq j$, this case is exactly the simple competitive learning algorithm.

For a compact region Ω of $\mathbb{R}^M$, let us introduce the definition of Voronoi tessellation associated with a family of vectors $y = (y_i, i = 1, \dots N) \in \Omega$.

DEFINITION 1. *For a given compact subset $\Omega \in \mathbb{R}^M$, the Voronoi tessellation $\Pi(y) = (\Pi(y)_i$, $i = 1, \dots, N)$ associated with a family of vectors $y_1, \dots, y_N$ is a partition of Ω given by*

$$\Pi(y)_i = \{x, \|y_i - x\| \leq \|y_j - x\|, j \neq i\}, \qquad i = 1, \dots, N.$$

Let us define a function g which is the leading term of the supermartingale difference given in the proof of Theorem 2.

$$g\left(y_1, y_2, \dots, y_N; w_1, w_2, \dots, w_N\right) = \sum_{i=1}^N (y_i - w_i) \cdot \left(\sum_k \int_{\Pi(y)_k} \Lambda(k,i)(x - y_i) f(x)\, dx \right).$$

g depends on the vectors $w = (w_1, w_2, \dots, w_N)$, $y = (y_1, y_2, \dots, y_N) \in \mathbb{R}^{M \times N}$. f is the density of the distribution P with support on a compact region Ω of $\mathbb{R}^M$, $\Pi(y) = (\Pi(y)_i,\ i = 1, \dots, N)$ is the Voronoi tessellation associated with $y = (y_1, \dots, y_N)$.

We define also:

$$\Theta := \{\text{the set of all Voronoi tessellations associated with } \{w_1(n), \dots, w_N(n)\}, \text{ for all } n\}.$$

For $y_1, \dots, y_N \in \mathbb{R}^M$, we use the convention that $y = (y_1, \dots, y_N) \in \Theta$ implies that there exists a Voronoi tessellation $\Pi(y)$ such that $\{\Pi(y)_i,\ i = 1, \dots N\} \in \Theta$.

Now we state the main theorem of this section.

THEOREM 2. *If there exists a unique point $(w_1, w_2, \dots, w_N) \in \mathbb{R}^{M \times N}$, such that*

$$g\left(y_1, \dots, y_N; w_1, \dots, w_N\right) \leq 0, \qquad \forall y \in \Theta, \tag{3}$$

where the equality holds, if and only if $y_i = w_i$, $i = 1, \dots, N$, *and*

$$\sum_n \eta(n) = \infty, \qquad \sum_n \eta^2(n) < \infty, \tag{4}$$

then we almost surely have

$$\lim_{n\to\infty} w_i(n) = w_i, \qquad i = 1, \dots, N.$$

Proof. We need to introduce some more notation. Let $\mathcal{F}_n$ be the sigma algebra generated by $\xi(k), k \le n$, $E(\zeta \mid \mathcal{F}_n)$ is the conditional expectation for the random variable ζ with respect to the sigma algebra $\mathcal{F}_n$. In terms of the proof of Theorem 5, below we see that

$$\sum_{i=1}^{N} \left[E\left(\|w_i(n+1) - w_i\|^2 \mid \mathcal{F}_n\right) - \|w_i(n) - w_i\|^2\right] \le \eta(n) g\left(w(n), w\right) + \eta(n)^2 g_1\left(w(n)\right), \tag{5}$$

for

$$g_1\left(w(n)\right) = \sum_{i=1}^{N} \sum_k \int_{\Pi(w(n)_k)} \Lambda(k,i) \|x - w_i(n)\|^2 f(x)\, dx$$

and $w(n) = (w_1(n), w_2(n), \dots, w(n))$. Since $g_1(w(n))$ is uniformly bounded by a constant A the inequality (5) thus, becomes

$$\begin{aligned} &\sum_{i=1}^{N} \left[E\left(\|w_i(n+1) - w_i\|^2 \mid \mathcal{F}_n\right) - \|w_i(n) - w_i\|^2\right] \\ &\quad \le \eta(n) g\left(w(n), w\right) + \eta(n)^2 A \\ &\quad \le \eta(n) g\left(w(n), w\right) + \eta(n)^2 A \left(1 + \sum_{i=1}^{N} \|w_i(n) - w_i\|^2\right). \end{aligned} \tag{6}$$

In terms of [14, Theorem 7.1, p. 43] together with Theorem 5 of the present paper, we arrive at the desired conclusions. ▮

Let us say a few words concerning condition (2). The fulfillment of condition (2) ensures that the learning algorithm moves downhill in the energy landscape, and so the uniqueness of the limit of the learning algorithm is true under condition (2). In Section 3, we will give a new family of learning rate parameters when condition (2) is violated, which is certainly the more interesting case.

For a one-dimensional input signal, i.e., $M = 1$, without loss of generality, we can assume that $a \le w_1(0) < w_2(0) < \cdots < w_N(0) \le b$ with $\Omega = [a, b]$. In this setting, we are able to simplify condition (2) in Theorem 2 due to the fact that the simple competitive learning does not change the order of weights $a \le w_1(n) < w_2(n) < \cdots < w_N(n) \le b$, $n \ge 1$.

Lemma 3. *If* $M = 1$, *then* $y_1, \dots, y_N \in \Theta$ *if and only if* $a \le y_1 < \cdots < y_N \le b$.

Proof. "$\Longrightarrow$." First note that if $w_1(0) < w_2(0) < \cdots < w_N(0)$, in the simple competitive learning, we still have $w_1(n) < w_2(n) < \cdots < w_N(n)$, for $n \ge 0$. Suppose that there is a Voronoi tessellation $\Pi \in \Theta$, then there exist

$$z_1 < z_2 < \cdots < z_N,$$

such that

$$\Pi_i = \left[\frac{z_{i-1} + z_i}{2}, \frac{z_i + z_{i+1}}{2}\right], \qquad i = 1, \dots, N,$$

here $(z_0 + z_1)/2 = a$ and $(z_{N+1} + z_N)/2 = b$. So $y \in \Theta$ implies that $a \le y_1 < y_2 < \cdots < y_N \le b$. "$\Longleftarrow$." Trivial. ▮

By combining Lemma 3 and Theorem 2, we have the following corollary. Three examples which explain the application of the next corollary are presented in Section 2.2.

COROLLARY 4. *If there exists a unique point $(w_1,\dots,w_N)\in\mathbb{R}^N$, such that the inequality*

$$\begin{aligned} g\left(y_1,\dots,y_N;\, w_1,\dots,w_N\right) &= \left(y_1-w_1\right)\sum_k \int_{(y_{k-1}+y_k)/2}^{(y_{k+1}+y_k)/2} \Lambda(k,1)\left(x-y_1\right) f(x)\,dx \\ &+ \sum_{i=2}^{N-1}\left(y_i-w_i\right)\sum_k \int_{(y_{k-1}+y_k)/2}^{(y_{k+1}+y_k)/2} \Lambda(k,i)\left(x-y_i\right) f(x)\,dx \\ &+ \left(y_N-w_N\right)\sum_k \int_{(y_{k-1}+y_k)/2}^{(y_{k+1}+y_k)/2} \Lambda(k,N)\left(x-y_N\right) f(x)\,dx \\ &< 0, \end{aligned}$$

holds for $a\le y_1<\cdots<y_N\le b$, except for $y_i=w_i$, $i=1,\dots,N$, and

$$\sum_n \eta(n)=\infty, \qquad \sum_n \eta^2(n)<\infty,$$

then we almost surely have

$$\lim_{n\to\infty} w_i(n)=w_i, \qquad i=1,\dots,N.$$

In the next theorem, we consider the convergence rate of the simple competitive learning. We prove that, under the conditions in Theorem 2, the algorithm will achieve the given accuracy within a finite number of updates.

We define

$$\tau(\epsilon)=\inf\left\{n,\, \|w_i(n)-w_i\|\le\epsilon,\ i=1,\dots,N\right\}$$

as the first time that the training error is less than ϵ.

THEOREM 5. *In the circumstances of Theorem 2, there exists a constant*

$$B(\epsilon)>0,$$

such that we have almost surely

$$\tau(\epsilon)<B(\epsilon).$$

PROOF. We find a negative bound for the difference:

$$\sum_{i=1}^N \left[E\left(\|w_i(n+1)-w_i\|^2 \mid \mathcal{F}_n\right)-\|w_i(n)-w_i\|^2\right],$$

and from it we get that $E(\|w_i(n+1)-w_i\|^2)$ is a supermartingale. According to the definition of the algorithm, we have

$$\begin{aligned} &E\left(\|w_i(n+1)-w_i\|^2 \mid \mathcal{F}_n\right)-\|w_i(n)-w_i\|^2 \\ &\quad= E\left(\|w_i(n+1)\|^2 \mid \mathcal{F}_n\right)-2w_i\cdot E\left(w_i(n+1)\mid\mathcal{F}_n\right)+\|w_i\|^2-\|w_i(n)-w_i\|^2 \\ &\quad= 2\eta(n)\left(w_i(n)-w_i\right)\cdot E\left(\left(\xi(n+1)-w_i(n)\right)I\left(w_i(n),\xi(n+1)\right)\mid\mathcal{F}_n\right) \\ &\qquad+\eta^2(n)E\left(\|\xi(n+1)-w_i(n)\|^2 I\left(w_i(n),\xi(n+1)\right)\mid\mathcal{F}_n\right). \end{aligned}$$

Since $w_i(n)$ and $\xi(n+1)$ are in the set

$$\left\{\|\xi(n+1)-w_i(n)\|\le\|\xi(n+1)-w_j(n)\|,\ j\ne i\right\},$$

if and only if

$$\xi(n+1) \in \Pi\,(w(n))_i\,,$$

for $w(n) = (w_i(n),\ i = 1, \dots, N)$, $w_i(n)$ and $\xi(n+1)$ are independent, we yield that

$$\begin{aligned}\eta(n)\,(w_i(n) - w_i)\cdot E\left((\xi(n+1) - w_i(n))\, I\,(w_i(n), \xi(n+1)) \mid \mathcal{F}_n\right)& \\ = \eta(n)\,(w_i(n) - w_i)\cdot \sum_k \int_{\Pi(w(n))_k} \Lambda(k,i)\,(x - w_i(n))\, f(x)\,dx& \qquad (7)\end{aligned}$$

and

$$\begin{aligned}\eta(n)^2 E\left(\|\xi(n+1) - w_i(n)\|^2\, I\,(w_i(n), \xi(n+1)) \mid \mathcal{F}_n\right)& \\ = \eta^2(n) \sum_k \int_{\Pi(w(n))_k} \Lambda(k,i)\,\|x - w_i(n)\|^2\, f(x)\,dx.& \qquad (8)\end{aligned}$$

Furthermore, if we replace the time n in equality (7) and (8) by the stopping time $\sigma_n := \tau(\epsilon) \wedge n = \min(n, \tau(\epsilon))$ all equalities hold. From the definition of the stopping time and condition (2) in Theorem 2, we see that

$$\begin{aligned}& g\,(w_1(\sigma_n), \dots, w_N(\sigma_n);\, w_1, \dots, w_N) \\ &\quad = \sum_{i=1}^{N} (w(\sigma_n)_i - w_i)\cdot \sum_k \int_{\Pi(w(\sigma_n))_k} \Lambda(k,i)\,(x - w(\sigma_n)_i)\, f(x)\,dx \qquad (9) \\ &\quad \le -h(\epsilon) < 0,\end{aligned}$$

for a number $h(\epsilon)$ depending only on ϵ. By condition (3) of Theorem 2, for n large enough, the sign of the term

$$\begin{aligned}&\eta(n) \sum_{i=1}^{N} (w(\sigma_n)_i - w_i)\cdot \sum_k \int_{\Pi(w(\sigma_n))_k} \Lambda(k,i)\,(x - w(\sigma_n)_i)\, f(x)\,dx \\ &\qquad + \eta^2(n) \sum_{i=1}^{N} \sum_k \int_{\Pi(w(\sigma_n))_k} \Lambda(k,i)\,\|x - w(\sigma_n)_i\|^2\, f(x)\,dx\end{aligned}$$

is determined by the sign of the following term

$$\begin{aligned}& g\,(w_1(\sigma_n), \dots, w_N(\sigma_n);\, w_1, \dots, w_N) \\ &\quad = \sum_{i=1}^{N} (w(\sigma_n)_i - w_i)\cdot \sum_k \int_{\Pi(w(\sigma_n))_k} \Lambda(k,i)\,(x - w(\sigma_n)_i)\, f(x)\,dx\end{aligned}$$

and so is negative, and we denote it $-h_1(\epsilon) < 0$. This explains the reason why we introduce the function g in Section 2. Without loss of generality, we assume that (9) is true for $n \ge 1$. We consider again the term

$$\sum_{i=1}^{N} \|w_i(\sigma_n) - w_i\|^2 + \sum_{k=1}^{\sigma_n - 1} h_1(\epsilon)\eta(k).$$

After repeating the same argument as before, we conclude that it is still a nonnegative supermartingale and so is

$$\sum_{i=1}^{N} \|w_i(\sigma_n) - w_i\|^2 + \sum_{k=1}^{\sigma_n - 1} h_1(\epsilon)\eta(k).$$

By the convergence of the supermartingale, the limit of

$$\sum_{i=1}^{N} \|w_i(\sigma_n) - w_i\|^2 + \sum_{k=1}^{\sigma_n - 1} h_1(\epsilon)\eta(k)$$

and

$$\sum_{i=1}^{N} \|w_i(\sigma_n) - w_i\|^2,$$

are both finite almost surely.

Thus,

$$\lim_{n\to\infty} \sigma_n = \lim_{n\to\infty} \tau(\epsilon) \wedge n < B,$$

almost surely for an integer B satisfying

$$\sum_{k=1}^{B} \eta(k)h_1(\epsilon) > N \max_{x,y\in\Omega} \|x-y\|^2 \geq \sum_{i=1}^{N} \|w_i(n) - w_i\|^2, \qquad \forall n,$$

which implies

$$\tau(\epsilon) < B$$

almost surely. Note that the random time $\tau(\epsilon)$ is bounded by a deterministic quantity B. ∎

Although that $w_i(n)$, $i = 1, \dots, N$ is a stochastic process, Theorem 5 asserts that within a finite and a deterministic time $B(\epsilon)$ $w_i(n)$, $i = 1, \dots, N$ will reach a given accuracy ϵ.

2.2. Examples

In this section, in order to show the applications of the theorems of the previous section, we consider three typical examples, in the sense that the first example takes into account the case when the input data set is discrete, the second and the third example consider the case when the input data set is continuously distributed according to the uniform distribution and the normal distribution, respectively. We consider only the case of simple competitive learning.

EXAMPLE 1. Suppose that $f(x) = \sum_{i=1}^{N} c_i \delta_{w_i}(x)$, with $\sum_{i=1}^{N} c_i = 1$ for $w_i \in [a,b] \subset \mathbb{R}^1$, $c_i > 0$, $i = 1, \dots, N$, and $w_1 < w_2 < \cdots < w_N$. Then we have that

$$\begin{aligned} g(y_1, \dots, y_N; w_1, \dots, w_N) = {} & -c_1 (y_1 - w_1)^2 I_{[a,(y_1+y_2)/2]}(w_1) \\ & - \sum_{i=2}^{N-1} c_i (y_i - w_i)^2 I_{[(y_{i-1}+y_i)/2,(y_i+y_{i+1})/2]}(w_i) \\ & - c_N (y_N - w_N)^2 I_{[(y_N + y_{N+1})/2,b]}(w_N). \end{aligned}$$

From the theorems of the previous section, we can conclude that

$$w = (w_1, \dots, w_N)$$

is the unique attracting point of the dynamics (1). The proof of Theorem 2 shows that the function g is the main contribution to the derivative of a Liapunov function. In fact, the quantity

$$\sum_{i=1}^{N} \left[E\left(\|w_i(n+1) - w_i\|^2 \mid \mathcal{F}_n\right)\right]$$

introduced in the proof can be considered to be the Liapunov function of the system. The difference appearing in the submartingale condition:

$$\sum_{i=1}^{N} \left[E\left(\|w_i(n+1) - w_i\|^2 \mid \mathcal{F}_n\right) - \|w_i(n) - w_i\|^2\right]$$

can be considered as a discretized derivative and is the sum of two terms. The one different from g vanishes. The points $(y_1, \dots, y_N)$ which make the function g equal to zero can be interpreted as the minima of this Liapunov function. Using this terminology one may say that the dynamics (1) will converge to the global minima $y_1 = w_1, \dots, y_N = w_N$ if the hypothesis of the Theorem 2 is satisfied. If there are many points for which the equality $g = 0$ is verified, then they may be seen as local minima which can trap the dynamics. The condition ensuring that there is a unique solution $(y_1, \dots, y_N)$ of the equation

$$g\left(y_1, \dots, y_N; w_1, \dots, w_N\right) = 0$$

is quite restrictive. In general, there are (infinitely) many solutions of it. Hence, the development of an algorithm to avoid the metastable states is of general importance, which is the content of the next section.

EXAMPLE 2. Suppose that $\xi(n)$ is uniformly distributed over the interval $[0,1]$. We are going to prove that $w_1 = 1/4$, $w_2 = 3/4$, and $g(y,w)$ is negative except for $y_1 = w_1$ and $y_2 = w_2$.

First note, that in this situation we have

$$g(y,w) = (y_1 - w_1)\int_0^{(y_1+y_2)/2} (x - y_1)\,dx + (y_2 - w_2)\int_{(y_1+y_2)/2}^{1} (x - y_2)\,dx.$$

Therefore,

$$\begin{aligned} g(y,w) &= (y_1 - w_1)\left(\int_0^{y_1} + \int_{y_1}^{(y_1+y_2)/2}\right)(x - y_1)\,dx + (y_2 - w_2)\left(\int_{(y_1+y_2)/2}^{y_2} + \int_{y_2}^{1}\right)(x - y_2)\,dx \\ &= (y_1 - w_1)\left(-\frac{1}{2}y_1^2 + \frac{1}{2}\frac{(y_1 - y_2)^2}{4}\right) + (y_2 - w_2)\left(-\frac{1}{2}\frac{(y_1 - y_2)^2}{4} + \frac{1}{2}(1 - y_2)^2\right). \end{aligned}$$

It is easy to check numerically (Figure 1) that $w_1 = 1/4$, $w_2 = 3/4$ is the unique point for $g(y,w) = 0$. Therefore, from Corollary 4 and Theorem 5 of the previous section, we have

$$\lim_{n\to\infty} w_1(n) = \frac{1}{4}, \qquad \lim_{n\to\infty} w_2(n) = \frac{3}{4},$$

and $\forall\,\epsilon > 0$, $\exists B(\epsilon) > 0$,

$$\tau(\epsilon) < B(\epsilon).$$

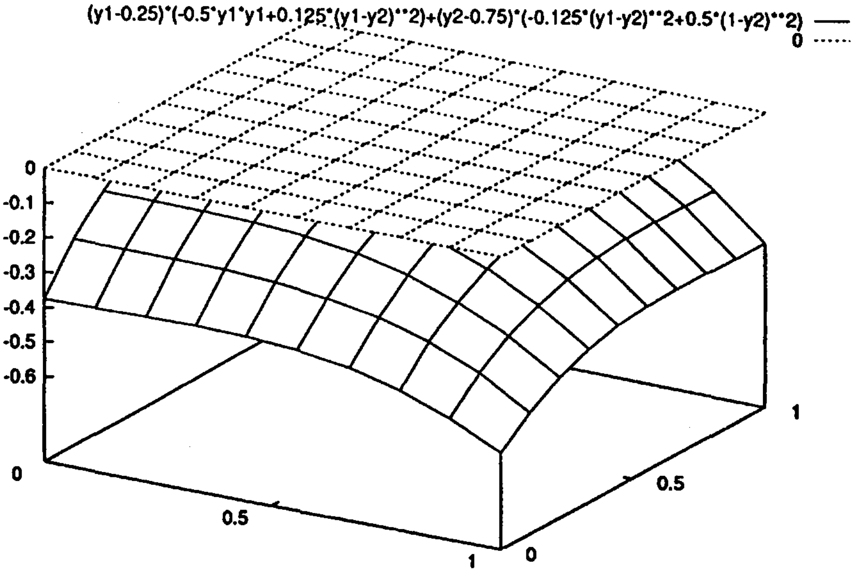

Figure 1. The function g defined in Example 2 of Section 2.

EXAMPLE 3. Suppose that $\xi(n)$ is distributed with density function

$$f(x) = \frac{1}{c} \exp\left(-\frac{x^2}{2}\right) I_{[-K,K]}(x),$$

the restriction of the normal distribution with mean 0 and variance 1 to $[-K, K]$, where $K = 2$ and $c = \int_{-K}^{K} e^{-(x^2/2)}\,dx$. It is natural to expect that $w_1 = -1/2c(1 - e^{-K^2/2}) = 0.18$ and $w_2 = 1/2c(1 - e^{-K^2/2})$.

Let

$$\begin{aligned}
\frac{g(y,w)}{c} &= (y_1 - w_1)\int_{-K}^{(y_1+y_2)/2} (x - y_1)\, e^{-x^2/2}\,dx + (y_2 - w_2)\int_{(y_1+y_2)/2}^{K} (x - y_2)e^{-x^2/2}\,dx \\
&= -(y_1 - w_1)\frac{1}{2}\, e^{-(y_1+y_2)^2/8} - (y_1 - w_1)y_1 \int_{-K}^{(y_1+y_2)/2} e^{-x^2/2}\,dx \\
&\quad + \frac{1}{2}\, e^{-K^2/2}\,(y_1 - y_2 - w_1 + w_2) + (y_2 - w_2)\frac{1}{2}\, e^{-(y_1+y_2)^2/8} \\
&\quad - (y_2 - w_2)y_2 \int_{(y_1+y_2)/2}^{K} e^{-x^2/2}\,dx.
\end{aligned}$$

It is easy to check numerically (see Figure 2) that the condition of Corollary 4 on the function g is not true, i.e., there are several points (y_1, y_2), $y_1 < y_2$ such that $g(y_1, y_2; w_1, w_2) = 0$.

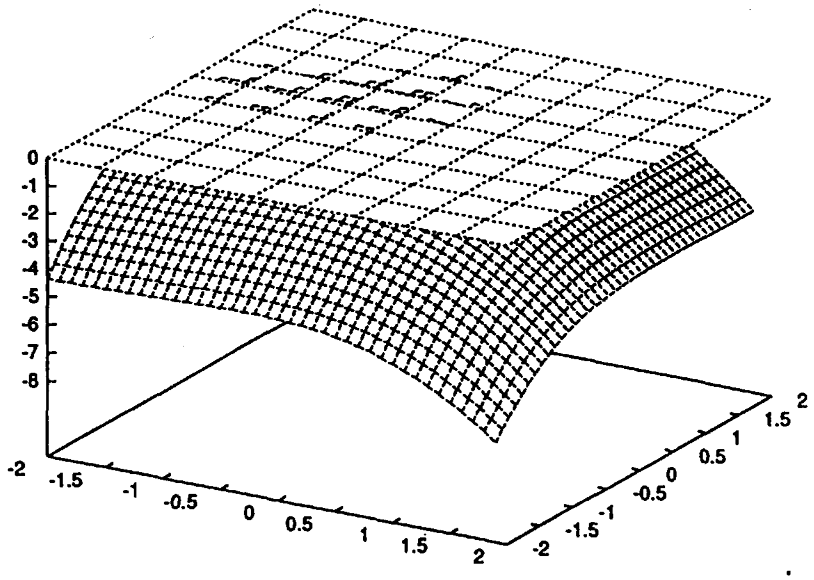

Figure 2. The function g defined in Example 3 of Section 2. Note that there are several points (y_1, y_2) with the property $g(y_1, y_2; w_1, w_2) = 0$.

3. A NEW FAMILY OF VLRPS

In Section 2, we developed a condition for the convergence of simple competitive learning with VLRPs. However, it is readily seen from our examples that, except for some special case (Example 2), the convergence of the algorithm will fail in general. On the other hand, all the algorithms similar to the simple competitive learning with VLRPs are in danger of getting caught in some local useless minima [4]. Hence, the problem of getting out of local minima is of general importance for the learning algorithms with VLRPs.

Essentially, a learning algorithm as we considered in Section 2 with VLRPs, can be written as

$$dX_t = \eta(t)\,(b(X_t)\,dt + \beta(t)\,dB_t)\,, \tag{10}$$

for $x_t \in \mathbb{R}^{M\times N}$, $b(\cdot)$ a measurable function on $\mathbb{R}^{M\times N}$, $t \in R^+$, and $\eta(t)$, the VLRPs with $\eta(0) \geq 0$, $\eta(t) \leq \eta(s)$ if $t \geq s$, $\beta(t) > 0$. Note that the discretized equation corresponding to (10) is

$$X_{n+1} = X_n + h\eta(n)b(X_n) + \beta(n)\sqrt{h}W_n, \tag{11}$$

where h is the step-size, W_n is normally distributed with zero mean, and covariance equal to the unit matrix I. For the purpose of finding a family of appropriate VLRPs in self-organizing Kohonen algorithm, equation (10) has been discussed in [15] from the Fokker-Planck equation point of view. In fact, in the field of neural networks, there are many learning algorithms developed with VLRPs and they are special cases of (10), for example, the network with Oja's rule [3], self-organizing Kohonen algorithm, the algorithm proposed in [18, p. 64], the dynamic link network [21,22], etc.

In this section, we first consider how to choose $\eta(t)$ and $\beta(t)$ so that X_t converges to the global minima of U if $b = -\text{grad } U$. It is proved in Theorem 6 below, that under the usual restriction (I) of Section 1 on $\eta(t)$, $\beta(t)$ should take the form (see Theorem 6)

$$\beta(t) \sim \frac{1}{\sqrt{\eta(t) \log \int_0^t \eta(u)\, du}}.$$

Note, that as $\eta(t) = c$ a constant independent of time, Theorem 6 reduces to the well-known results of simulated annealing [16].

Second, if the signal is not separable from the noise, this means that in the equation (10) we require $\beta(t) = 1, \forall t$. It is shown in Theorem 7 below, that if the family of VLRPs $\eta(t)$ satisfies an ODE, the solution of which is between $1/\log t$ and $1/\sqrt{\log t}$, then X_t will converge to the global minima with probability one. It is worthwhile to point out that this family of VLRPs already does not satisfy the restriction (II) of Section 1. We believe that the discovery of this section is of general importance, also for some well-known statistical algorithms such as Robbis-Monro procedure and Kiefer-Wolfowitz procedure, which have been intensively studied in the statistics (see [10,11,13,14]) and take the form of (10). For the neural network applications of these algorithms we refer the reader to [23, Chapter 2].

3.1. The General Case

In this section, we consider equation (10):

$$dX_t = \eta(t)\left(b(X_t)\, dt + \beta(t)\, dB_t\right).$$

In order to develop a new learning rate for ensuring the convergence of the algorithm to the global minima, we apply the results of simulated annealing to our case [16]. However, simulated annealing corresponds to the case in which the dynamics without noise is homogeneous, namely $\eta(t)$ is a constant independent of time t, and the noise goes to zero as the system evolves. This requires that in (10), the coefficient in front of b should be independent of t, while there is still a vanishing rate before the Brownian motion B_t. Fortunately, after taking another time scaling, we are able to remove the vanishing term in front of the drift term b, and keep the second term of the noise as a standard Brownian motion because of the self-similarity property of the Brownian motion. Furthermore, there is still a vanishing rate multiplying the Brownian motion.

Before going to more general cases, we show here first an example, in order to explain our general ideas above.

EXAMPLE 4. Take $\eta(t) = 1/t$, $\beta(t) = 1$, $M = N = 1$ in equation (10). Note that in this setting, the conditions (I) and (II) of Section 1 are fulfilled for the choice of $\eta(t)$. Now the dynamics (10) reads

$$dX_t = \frac{1}{t}\left(b(X_t)\, dt + dB_t\right).$$

In order to change the time scaling of the above dynamics, let us make an change of the time scale:

$$s = \log t = \int_0^t \eta(u)\, du, \tag{12}$$

or

$$t = \exp(s),$$

and $Y_s = X_{e^s}$.

Then

$$dX_{e^s} = dY_s = \frac{1}{e^s} b(Y_s) e^s\, ds + \frac{1}{e^s}\, dB_{e^s}. \tag{13}$$

From the self-similarity property of the Brownian motion, we know that

$$e^{-s/2}\, dB_{e^s} \sim N(0,\, ds).$$

So we introduce a new time scaling s and write $d\tilde{B}_s = e^{-s/2} dB_{e^s}$, $\tilde{B}_s$ is again a standard Brownian motion. Now (3) can be rewritten as

$$dY_s = b(Y_s)\, ds + e^{-s/2}\, d\tilde{B}_s. \tag{14}$$

The relation (12) between the time t and s tells us that if s goes to infinity, then t goes to infinity also, and vice versa. So if we know the limit behavior of Y_s, we know the limit behavior of X_t as well. From the general results of simulated annealing [16,24,25], we know that in the case of equation (14), Y_s will have positive probability to stay at any local minimum since the noise vanishes too fast, at a rate of $\exp(-s/2)$. In order to ensure that X_t is not trapped in some local minima, we should slow down the decreasing rate of the noise. For this example, a correct choice is (see Theorem 6)

$$dZ_t = \eta(t) \left[b(Z_t)\, dt + \frac{\gamma}{\sqrt{\eta(t)} \log\log t}\, dB_t \right], \tag{15}$$

for a constant γ, which as in simulated annealing is problem dependent.

In [13,14], under the restriction of

$$\int_0^\infty \eta(u)\, du = \infty, \qquad \int_0^\infty \eta(u)^2 \beta(u)^2\, du < \infty, \tag{16}$$

for the stochastic differential equation (I) and (II) of Section 1 is a special case of (16),

$$dX_t = \eta(t) b(X_t)\, dt + \eta(t)\beta(t)\, dB_t,$$

the convergence of the solution to the set of attractors (no global minima!) of the above dynamics is proved. However, we note that in equation (15), the VLRPs $\eta(t) = 1/t$, $\beta(t)\eta(t) = 1/\sqrt{t \log\log t}$ with

$$\int_0^\infty \eta(u)\, du = \infty, \qquad \int_0^\infty \eta(u)^2 \beta(u)^2\, du = \int_0^\infty \frac{1}{u \log\log u}\, du = \infty,$$

already violate the usual restriction (16) found in stochastic approximation theory.

In general, we have the following result for $b(x) = -\text{grad}\ U(x)$ for a function U defined on Ω (see Remark 2).

THEOREM 6. *Suppose that*

$$\lim_{t\to\infty} \int_0^t \eta(u)\, du = \infty \tag{17}$$

and

$$dZ_t = \eta(t)\left[b(Z_t)\,dt + \frac{\gamma}{\sqrt{\eta(t)\log\int_0^t \eta(u)du}}\,dB_t\right], \tag{18}$$

where $Z_t \in \Omega$, a compact subset of $\mathbb{R}^{M\times N}$, b is a measurable function on $C^1(\Omega)$, and B_t is the $M \times N$-dimensional Brownian motion. Then there exists a constant γ_0, and a set $A \subset \Omega$ such that as $\gamma > \gamma_0$, we have

$$\lim_{t\to\infty} P(Z_t \in A) = 1,$$

where A is the set of global minima of U.

PROOF. Let

$$s = s(t) = \int_0^t \eta(u)\,du \tag{19}$$

denote its inverse function as $t = t(s)$. Define

$$Y_s = Z_t = Z_{t(s)}.$$

Then the equation (18) becomes

$$dY_s = b(Y_s)\,ds + \frac{\gamma\sqrt{\eta(t)}}{\sqrt{\log s}}\,dB_t. \tag{20}$$

In terms of the self-similarity property of the Brownian motion and $\eta(t)dt = ds$, we derive that

$$d\tilde{B}_s := \sqrt{\eta(t(s))}\,dB_{t(s)} \sim N(0, ds\cdot I),$$

where I is the $(M \times N) \times (M \times N)$ unit matrix. Hence, $\tilde{B}_s$ is still a standard Brownian motion on $\mathbb{R}^{M\times N}$. Now (18) becomes

$$dY_s = b(Y_s)\,ds + \frac{\gamma}{\sqrt{\log s}}\,d\tilde{B}_s. \tag{21}$$

From the condition of the present theorem, we see that

$$s(t) \to \infty, \qquad \text{as } t \to \infty,$$

and

$$t(s) \to \infty, \qquad \text{as } s \to \infty.$$

Therefore, we have

$$\lim_{t\to\infty} P(Z_t \in F) = \lim_{s\to\infty} P(Y_s \in F),$$

for any measurable subset F of $\mathbb{R}^M$.

By theorems of [16], we deduce that there is a positive constant γ_0 such that as $\gamma > \gamma_0$,

$$\lim_{t\to\infty} P(Z_t \in A) = \lim_{s\to\infty} P(Y_s \in A),$$

where A is the set of the minima of U as $b(x) = -\text{grad}\,U(x)$. ∎

REMARK 1. In Theorem 6, γ_0 could be (roughly) chosen to equal to

$$\sqrt{2\left(\sup_{x\in\Omega} U(x) - \inf_{x\in\Omega} U(x)\right)}.$$

REMARK 2. If there is no energy function U for the dynamics, the action functional defined by

$$A(x,y) = \inf_{\phi} \left\{ S_{0,T}(\phi); S_{0,T}(\phi) = \frac{1}{2} \int_0^T \|\phi'_t - b(\phi_t)\|^2 \, dt, \right.$$
$$\left. \phi \in C_{[0,T]}\left(\mathbb{R}^M\right), \ \phi_0 = x, \ \phi_T = y, \ \forall T \geq 0 \right\},$$

could be used to replace U and

$$\gamma_0 = \sqrt{2 \sup_{x,y \in \Omega} A(x,y)}.$$

Similar results as in the above theorem are still true, see [16,24].

REMARK 3. Our approach also yields a conclusion which is already noted in [15, p. 259]. When $b(x) = -x$ in equation (10), it is pointed out in [15, p. 259], that the second condition (II) of Section 1, i.e.,

$$\int_0^\infty \eta^2(u) \, du < \infty$$

can be replaced by a much weaker condition

$$\lim_{t \to \infty} \eta(t) = 0,$$

and the conditions

$$\int_0^\infty \eta(u) du = \infty, \qquad \lim_{t \to \infty} \eta(t) = 0,$$

are necessary and sufficient for X_t to converge to 0. In fact, our approach also rigorously yields this result. Consider the equation of X_t

$$dX_t = \eta(t) \left[b(X_t) \, dt + dB_t \right].$$

After taking the new time scaling s (see the proof of Theorem 6), we yield that

$$dY_s = b(Y_s) \, ds + \sqrt{\eta(s)} \, d\tilde{B}_s,$$

if $\eta(t) = \eta(t(s)) \to 0$ and $U(x)$ only one minimum, say x_0 (the case considered in [15] $x_0 = 0$), we know that $X_t \to x_0$, a.s. This proves the sufficiency. The necessary condition is obvious since if $\eta(t)$ does not go to zero, Y_s will certainly not stay at x_0 at all.

REMARK 4. We can of course choose a family of VRLPs decreasing more slowly, and at the same time ensure that the conclusions of Theorem 6 are still true. For example, if we set

$$\beta(t) = \frac{\gamma}{\sqrt{\eta(t) \log \log \int_0^t \eta(u) \, du}},$$

then we still have the conclusions of Theorem 6.

3.2. A Special Case

In some situations, it is not possible to separate the drift term b from the Brownian motion B_t. And sometimes the data sent as an input to the network is noise-contaminated also. This is equivalent to asking if there exists a family of $\eta(t)$ such that X_t converges to the global minima of U, where X_t is the solution of

$$dX_t = \eta(t) \left(b(X_t) \, dt + dB_t \right).$$

From Theorem 6, we know that above requirement is equivalent to say that for $t \geq 1$

$$\eta(t) \log \int_0^t \eta(u)\, du = \gamma^2 \tag{22}$$

or

$$\log \int_0^t \eta(u)\, du = \frac{\gamma^2}{\eta(t)}.$$

Differentiating on both sides of the equation above, we have

$$\frac{\eta(t)}{\int_0^t \eta(u)\, du} = -\frac{\eta'(t)\gamma^2}{\eta(t)^2}$$

or

$$\eta'(t) = -\frac{\eta^3(t)}{\gamma^2 \int_0^t \eta(u)\, du}. \tag{23}$$

If we are able to solve the above equation and prove that its solution satisfies the conditions of Theorem 6, we obtain a family of new VLRPs $\eta(t)$. $\eta(t)$ could be easily computed numerically (Figure 3) and we have the following estimate.

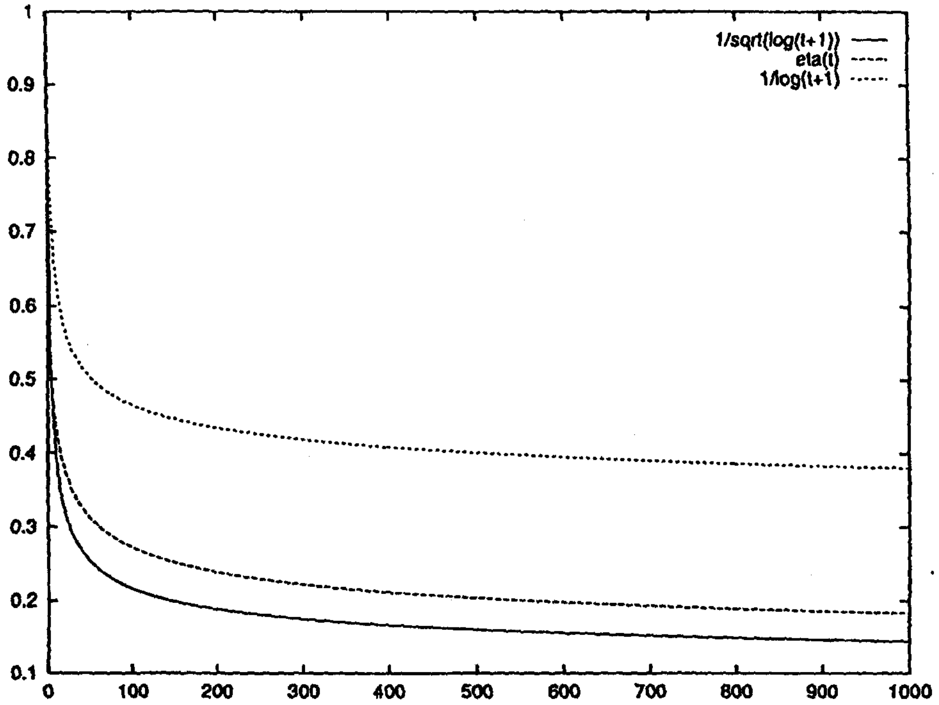

Figure 3. The function $\eta(t)$ with $\gamma = 1$ defined in Theorem 7 of Section 3.

In terms of the nonincreasing property of $\eta(t)$, we have

$$-\frac{\eta^2(t)}{\gamma^2 t} \leq \eta'(t) \leq -\frac{\eta^3(t)}{\eta(1)\gamma^2 t},$$

for $t \geq 1$, which implies that

$$\frac{\gamma^2 \eta(1)}{\eta(1) \log t + \gamma^2} \leq \eta(t) \leq \frac{\gamma \eta(1)}{\sqrt{\eta^2(1) \log t + \gamma^2}},$$

for $\gamma \geq \gamma_0$. This also proves that the condition (17) in Theorem 6, for $\eta(t)$ is fulfilled.

By combining Theorem 6 and all conclusions above, we now come to the main theorem of the present paper. We say that a family of VLRPs is optimal if it guarantees the learning algorithm to converge to the global minima of the energy function.

THEOREM 7. *Let us consider the stochastic differential equation*

$$dX_t = \eta(t)\,(b(X_t)\,dt + dB_t)\,.$$

A family of optimal VLRPs in the learning algorithm with VLRPs is the unique solution of the equation

$$\eta'(t) = -\frac{\eta^3(t)}{\gamma^2 \int_0^t \eta(u)\,du}, \qquad t \ge 1, \tag{24}$$

with $\eta(1) = \gamma^2/\log\int_0^1 \eta(u)\,du$ *(see Figure 3).* $\eta(t)$ *is bounded from below and above:*

$$\frac{\gamma^2\eta(1)}{\eta(1)\log t + \gamma^2} \le \eta(t) \le \frac{\gamma\eta(1)}{\sqrt{\eta(1)^2\log t + \gamma^2}}, \qquad t \ge 1,$$

for some positive constants $\gamma \ge \gamma_0$*, where* γ_0 *is defined as in Theorem 6.*

PROOF. It suffices for us to prove the uniqueness of the solution equation (22) and the differentiability of the solution. We use the contraction mapping theorem. For this purpose we write

$$S\,[1+(n-1)\delta,\ 1+n\delta] = \{\eta \in C\,[1+(n-1)\delta,\ 1+n\delta] \text{ and } \eta(t) \ge 0\}\,, \tag{25}$$

which is a closed subset of $C[1+(n-1)\delta,\ 1+n\delta]$ and δ will be specified later. Set

$$T_0(\eta) = \frac{\gamma^2}{\log(\int_1^t \eta(u)\,du + c_1)}, \tag{26}$$

a mapping from $S[1, 1+\delta]$ onto itself with $c_1 = \int_0^1 \eta(u)\,du > 0$. For $\eta_1, \eta_2 \in S[1, 1+\delta]$ with $\eta_1(u) = \eta_2(u) = \eta(u)$, $0 \le u \le 1$ we have

$$\begin{aligned} \|T_0(\eta_1) - T_0(\eta_2)\| &= \max_{t\in[1,1+\delta]} \left| \frac{\gamma^2}{\log\left(\int_1^t \eta_2(u)\,du + c_1\right)} - \frac{\gamma^2}{\log\left(\int_1^t \eta_1(u)\,du + c_1\right)} \right| \\ &\le \frac{\gamma^2}{(\log c_1)^2} \max_{t\in[1,1+\delta]} \left| \log\left(\int_1^t \eta_2(u)\,du + c_1\right) - \log\left(\int_1^t \eta_1(u)\,du + c_1\right) \right| \\ &= \frac{\gamma^2}{(\log c_1)^2} \max_{t\in[1,1+\delta]} \left| \log\left(1 + \frac{\int_1^t |\eta_2(u) - \eta_1(u)|\,du}{\int_1^t \eta_1(u)\,du + c_1}\right) \right|. \end{aligned} \tag{27}$$

From the basic inequality $\log(1+x) \le x$ for $x \ge 0$, we deduce that

$$\|T_0(\eta_1) - T_0(\eta_2)\| \le \frac{\gamma^2}{(\log c_1)^2} \frac{\int_1^{1+\delta} \|\eta_1 - \eta_2\|\,du}{c_1}. \tag{28}$$

Hence, as $\delta < (c_1(\log c_1)^2)/\gamma^2$ the mapping T_0 is a contraction mapping, and so on the space $S[1, 1+\delta]$ there exists a unique ξ such that it satisfies (22).

Next, we use induction for the proof of the existence and uniqueness of η on the time interval $[1+n\delta, 1+(n+1)\delta]$. Assume that, we have proved there exists a unique solution η on time interval $[1, 1+n\delta]$ denoting it as $(\xi(t))_{1\le t\le 1+n\delta}$. Define a mapping $T_n(\eta)$ for $\eta \in S[1+n\delta, 1+(n+1)\delta]$ by

$$T_n(\eta)(t) = \frac{\gamma^2}{\log\left(\int_1^t \eta(u)\,du + c_1\right)}, \tag{29}$$

where $\eta(u) = \xi(u)$ for $0 \le u \le n\delta$. By repeating the above arguments for $n = 0$, we conclude that T_n is again a contraction mapping in the complete space $S[1+n\delta, 1+(n+1)\delta]$. We assert the existence and uniqueness of the solution of equation (22) writing it as η.

Now, we prove that $\eta(t)$ fulfills (24). In fact, from equation (22) we see that $\eta(t) > 0$ and $\eta(t)$ is differentiable with $\eta'(t) < 0 \, \forall t \geq 1$. Differentiating on both sides of (22) with respect to t we yield equation (24). ∎

Next, we present numerical simulations for a simple model. The reason for us to consider this simple model here is that we can find γ_0 exactly.

Example 5. Let $U(x) = x^4 + x^3 - 4x^2 + x$ (see Figure 4). We have a numerical comparison of the following three kinds of dynamics.

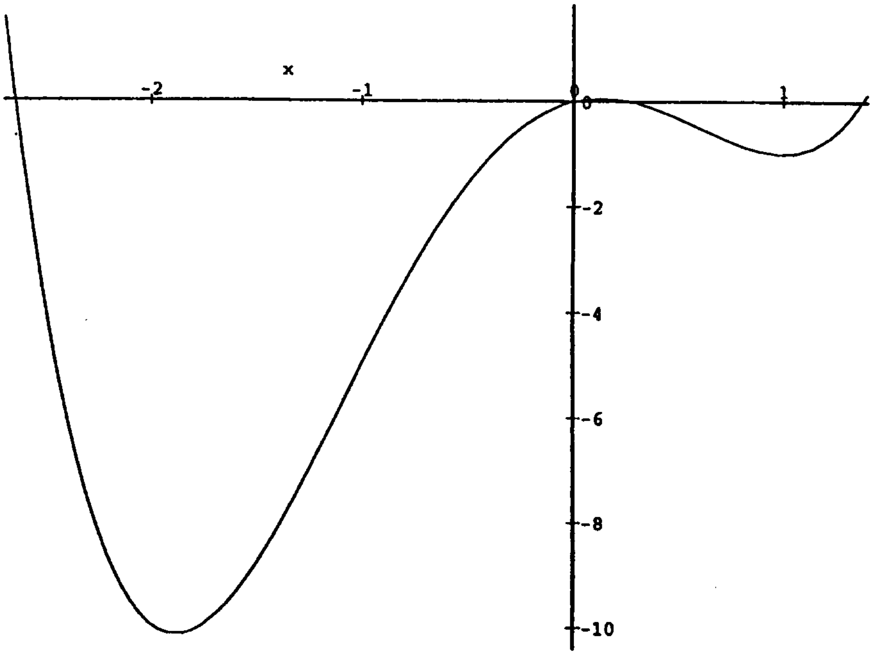

Figure 4. The potential function $U(x) = x^4 + x^3 - 4x^2 + x$. There are two minima, one is at $x = 1$ (a local minimum) and another is at $x = (-7 - \sqrt{65})/8 = -1.81$ (global minimum).

Thus, we consider the algorithm with the VLRPs of Theorem 7,

$$du_t = -\eta(t)\left(U'(x_t)\,dt + dB_t\right) \tag{30}$$

the algorithm of simulated annealing

$$dv(t) = -U'(v_t)\,dt + \frac{\gamma}{\sqrt{\log(t+2)}}\,dB_t, \tag{31}$$

and the algorithm with VLRPs of $1/t$

$$dw_t = -\frac{1}{t}\left(U'(w_t)\,dt + dB_t\right). \tag{32}$$

We discretized them with time step $h = 0.01$ (see equation (11)) and with initial state $u_0^{(j)} = v_0^{(j)} = w_0^{(j)} = 0.1j - 1$, where $j = 0, \ldots, 20$ namely we carry out 21 simulations with initial state from $[-1, 1]$ for dynamics u_t, v_t, and w_t. For each given j after 50000 iterations we get a solution $u(j)$, $v(j)$, and $w(j)$ corresponding to dynamics (30),(31), and (32), respectively.

Finally, we have

$$u = \frac{\sum_{j=1}^{21} u(j)}{21} = -1.76, \qquad v = \frac{\sum_{j=1}^{21} v(j)}{21} = -1.88, \qquad w = \frac{\sum_{j=1}^{21} w(j)}{21} = -0.36,$$

Table 1. Numerical results of three algorithms (alg.) for initial states from −1. to 1. Note, that only starting from 0.6 the algorithm $u(j)$ fails to arrive to the global minima.

Initial state	−1.00	−0.90	−0.80	−0.70	−0.60	−0.50	−0.40	
Alg. 1$u(j)$	−1.87	−1.88	−2.00	−1.95	−1.73	−1.91	−1.83	
Alg. 2$v(j)$	−1.87	−1.79	−2.00	−1.90	−1.83	−1.85	−1.77	
Alg. 3$w(j)$	−1.80	−1.73	−1.78	−1.58	−1.55	−1.58	−1.02	
Initial state	−0.30	−0.20	−0.10	0.00	0.10	0.20	0.30	
Alg. 1$u(j)$	−1.81	−2.14	−1.85	−1.93	−1.91	−1.88	−1.77	
Alg. 2$v(j)$	−1.78	−2.12	−1.87	−1.85	−1.93	−1.88	−1.74	
Alg. 3$w(j)$	−1.27	−1.42	−0.50	−0.02	0.21	0.21	0.25	
Initial state	0.40	0.50	0.60	0.70	0.80	0.90	1.00	mean value
Alg. 1$u(j)$	−1.90	−1.83	0.97*	−1.97	−1.97	−1.92	−1.82	−1.76
Alg. 2$v(j)$	−1.82	−1.90	−1.92	−1.92	−1.95	−1.87	−1.84	−1.88
Alg. 3$w(j)$	0.68	0.60	0.84	0.97	1.03	0.97	1.02	−0.36

the average of dynamics (30), (31), and (32) over 21 different initial states. Note that the exact global minima is at $x = -1.81$. The parameter γ is set to 2 (we refer the reader to [16] for an explanation of the choice of this value, γ_0 in Remark 1 is a rather rough choice).

As we expected, the dynamics (32) will stay at the local minima with highest probability among dynamics (30), (31), and (32). Dynamics (30) and (31) are more likely to go to the global minima. For the dynamics (18) all 21 simulations are successful in finding the global minima. For dynamics (17) 20 simulations are successful in finding the global minima but one fails. Our numerical results here confirm our theoretical approach.

Finally, a comment should be made about the practical use of the theory presented in this section. Typically, there are two ways to associate the dynamics (10) with a learning algorithm such as self-organizing Kohonen algorithm, Hebb-type learning, etc. One is to consider the learning algorithm:

$$\frac{dx_t}{dt} = \eta(t)b(x_t). \tag{33}$$

If we suppose that some stochastic noises are contained in the model, the simplest assumption of it is that now the dynamics (33) takes the form (10). Note that our Theorems 6 and 7, are proved without any restriction on b except that $b \in C^1(\Omega)$ (see Remark 2 and [16]), and so it is general enough to cover learning algorithms developed in neural networks[1]. In this situation, for avoiding local minima, our approach suggests that it is more reasonable to use the family of vanishing learning rate parameters in Theorem 4 than the one of order $1/t^\alpha$, $1/2 < \alpha \le 1$. Another way is that the term $\eta(t)B_t$ might be added artificially, following the usual logic of the "annealing" scheme, in order to force the dynamics to jump around until it eventually "settles" near a global minimum. For example, for the simple competitive learning defined by equation (1) let $b(x) = (E\bar{I}(x, \xi(n+1))(\xi_j(n+1) - x_{ij}), i = 1, \dots, N, j = 1, \dots, M)$, where E is the expectation with respect to ξ, by adding a noise term $\eta(t)B_t$ to the learning algorithm, we assert that the algorithm will reach a global minimum.

4. CONCLUSIONS

Basically, we consider two questions in the present paper. First, a convergence theorem for Kohonen self-organizing map is presented. The same result for the simple competitive learning follows as corollary. Secondly, we rigorously derive a new family of vanishing learning rate parameters for a useful class of learning algorithm.

[1] [5] It is proved that there is no function U for self-organizing Kohonen algorithm with the property $b = -\text{grad } U$.

Global optimization of learning in neural networks is currently an important subject. How can one be sure that the learning network reaches the optimal state, i.e., the global minimum of some error criterion, and does not get stuck in a local minimum? A well known strategy to find the global minimum and not just a local minimum is simulated annealing [16], a noise parameter, say temperature, is cooled down slowly. More specifically, we consider the following stochastic differential equation (or Langevin equation)

$$dX_t = -\text{grad } U(X_t)\, dt + \alpha(t)\, dB_t, \tag{34}$$

and when

$$\alpha(t) = \frac{\gamma}{\sqrt{\log(t+2)}}, \tag{35}$$

we have

$$\lim_{t\to\infty} P\left(X_t \in A\right) = 1,$$

where A is the set of global minima of U, and γ is a constant depending on U.

Learning in neural networks such as self-organizing Kohonen algorithm, Hebb learning, etc., are also a stochastic process. At each learning step, a training pattern is drawn at random from the environment (the total set of training patterns) and presented to the network. A large learning parameter leads to large fluctuations in the synaptic weight of the network. So, in a way, the learning parameter can be viewed as a noise parameter akin to the temperature in simulated annealing. A typical case of such learning algorithms (see, final chapter, of previous section) is

$$dY_t = \eta(t)\left(b(Y_t)\, dt + \beta(t)\, dB_t\right), \tag{36}$$

a dynamics studied in stochastic approximation theory for many years. Note that when $b = -\text{grad } U$, $\eta(t) = 1$, and $\alpha(t) = \beta(t)$ we have $X_t = Y_t$, and thus, the case for simulated annealing is just a special case of (36).

In the present paper, we derive a family of vanishing learning rate parameters based upon a rigorous analysis on (36) and our previous results of simulated annealing in [16]. The new family of vanishing learning rate parameters satisfy the following condition

$$\begin{gathered}\int_0^\infty \eta(u)\, du = \infty, \\ \beta(t) = \frac{\gamma}{\sqrt{\eta(t) \int_0^t \eta(u)\, du}},\end{gathered} \tag{37}$$

which in general violates the condition (10) found in stochastic approximation theory. Again we want to point out here that when $\eta(u) = 1$, the rate (22) found in simulated annealing algorithm defined by (34) is exactly a special case of our results here.

Finally, we like to comment on further possible developments of our results here. Obviously a case to case and systematic numerical simulations for algorithms developed in neural networks with VLRPs in Theorem 6 and Theorem 7, are quite interesting and is one of our further topics. Theoretically simulated annealing of form (34) has been well studied [16] and on the other hand stochastic approximation theory taking into account the dynamics (36) has developed into a mature field already. In particular, many estimates on convergence rate (in neural networks, convergence rate is called learning error and generalization error) for both algorithms have been established already. We believe that the method developed in this paper serves as a bridge between these two fields and will help us to understand more deeply the behavior of learning algorithms in neural networks and may provide a theoretical basis for the design of practical algorithms that lead to global optimization of learning in neural networks.

REFERENCES

1. S. Albeverio, N. Krüger and B. Tirozzi, An extension of Kohonen phonetic maps for speech recognition, *Mathl. Comput. Modelling* (to appear).
2. D. Amit, *Modeling Brain Function*, Cambridge Univ. Press, (1989).
3. J.A. Hertz, A. Krogh and R. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, (1991).
4. E. Erwin, K. Obermayer and K. Schulten, Self-organizing maps: Stationary states, metastability and convergence rate, *Biol. Cybern.* **67**, 35–45, (1992).
5. E. Erwin, K. Obermayer and K. Schulten, Self-organizing maps: Ordering, convergence properties and energy functions, *Biol. Cybern.* **67**, 47–55, (1992).
6. C. Bouton and G. Pagès, Self-organization and convergence of the one-dimensional Kohonen algorithm with non uniformly distributed stimuli, *Stoch. Proc. Appl.* **47**, 249–274, (1993).
7. C. Bouton and G. Pagès, Convergence in distribution of the one-dimensional Kohonen algorithms when the stimuli are not uniform, *Adv. Appl. Prob.* **26**, 80–103, (1994).
8. M. Cottrell and J.C. Fort, Etude d'un algorithme d'auto-organisation, *Ann. Inst. H. Poincaré* **23**, 1–20, (1986).
9. T. Martinetz and K. Schulten, Topology representing networks, *Neural network* **7**, 507–522, (1994).
10. A. Benveniste, M. Métivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, (1990).
11. P. Hall and C.C. Heyde, *Martingale Limit Theory and its Application*, Academic Press, New York, (1980).
12. T. Kohonen, *Self-Organization and Associative Memory*, 3rd Edition, Springer-Verlag, Berlin, (1989).
13. H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, (1978).
14. M.B. Nevel'son and R.Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*, Translation of Math. Monograph 47, Amer. Math. Soc., Providence, RI, (1976).
15. H. Ritter, T. Martinetz and K. Schulten, *Neural Computation and Self-Organizing Maps*, Addison-Wesley, Reading, MA, (1992).
16. S. Albeverio, J.F. Feng and M.P. Qian, Role of noises in neural networks, *Phys. Rev. E* **52**, 6593–6606, (1995).
17. J.F. Feng, H. Pan and V.P. Roychowdhury, On neurodynamics with limiter function and Linsker's developmental model, *Neural Computation* **8**, 1003–1019, (1996).
18. S. Amari, Information geometry of the EM and em algorithms for neural network, *Neural Networks* **8**, 1379–1408, (1995).
19. G.J. Goodhill, Topography and ocular dominance:a model exploring positive correlations, *Bio. Cyber.* **69**, 109–118, (1993).
20. G.J. Goodhill and D.J. Willshaw, Elastic net model of ocular dominance: Overall stripe pattern and monocular deprivation, *Neural Computation* **6**, 615–621, (1994).
21. J.F. Feng and B. Tirozzi, A discrete version of the dynamic link network, *Neurocomputing* **14** (to appear).
22. M. Lades, J.C. Vorbrüggen, J. Buhrmann, J. Lange, C. von der Malsburg, R.P. Würtz and W. Konen, Distortion invariant object recognition in the dynamic link architecture, *IEEE Transactions on Computers* **42**, 300–311, (1993).
23. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, (1995).
24. J.F. Feng and M.P. Qian, Two-stage annealing in retrieving memories I, In *Probability and Statistics*, (Edited by A. Badrikian, P.-A. Meyer and J.-A. Yan), pp. 149–176, World Scientific, Singapore, (1993).
25. M.I. Freidlin and A.D. Wentzell, *Random Perturbations of Dynamic System*, Springer-Verlag, Berlin, (1984).
26. S.Y. Chow and H. Teicher, *Probability Theory*, Springer-Verlag, New York, (1988).