A Class of Competitive Learning Models which Avoids Neuron Underutilization Problem

Clifford Sze-Tsan Choy and Wan-Chi Siu, *Senior Member, IEEE*

***Abstract*— In this paper, we study a qualitative property of a class of competitive learning (CL) models, which is called the multiplicatively biased competitive learning (MBCL) model, namely that it avoids neuron underutilization with probability one as time goes to infinity. In the MBCL, the competition among neurons is biased by a multiplicative term, while only one weight vector is updated per learning step. This is of practical interest since its instances have computational complexities among the lowest in existing CL models. In addition, in applications like classification, vector quantizer design and probability density function estimation, a necessary condition for optimal performance is to avoid neuron underutilization. Hence, it is possible to define instances of MBCL to achieve optimal performance in these applications.**

***Index Terms*— Multiplicatively biased competitive learning, neuron underutilization problem, vector quantization.**

I. Introduction

COMPETITIVE learning (CL) model is a class of unsupervised[1] learning models, which is characterized by competitions among its units. To confine our scope of discussions in this paper, we define a general CL model as follows. A CL model has N units, in which the ith unit has an associated *state* S_i. Then, the state of a CL model is denoted by $S = \{S_1, \cdots, S_N, \Gamma\}$, which is the collection of states of N units as well as a global state Γ representing information common to all units (e.g., topological relationship in the self-organizing map (SOM) [2]). At each learning step, an input vector $\mathbf{x}$ is presented to the model, which can be considered as a sample from an *input space* denoted by Ω. Units compete among each other according to their relative fitness to the input vector $\mathbf{x}$, and a unique *winner* is selected, which is necessarily the one with the highest relative fitness. The ith unit is assigned a *competition score* $\gamma_i(\mathbf{x}, S, t)$ to quantify its relative fitness to the input vector $\mathbf{x}$ at the current state S and at time t, such that the winner is necessarily the one with the highest score. The winner is then *rewarded* most, although losers may be rewarded as well to a lesser extent. In saying that the ith unit is rewarded, we mean that its state is modified such that its relative fitness to the input vector $\mathbf{x}$ is increased.

Manuscript received August 5, 1997; revised April 21, 1998. This work was supported by the Research Grant Council of the Hong Kong Special Administrative Region under Grant HKP152/92E (PolyU340/939).

The authors are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

Publisher Item Identifier S 1045-9227(98)08350-7.

[1] Although there are supervised CL models under the name learning vector quantization (LVQ) studied by Kohonen [1], we only consider unsupervised CL models in this paper.

With the concept of winner in the CL model, one can think of the input space Ω as being partitioned exhaustively into N disjoint regions $\{\Omega_1, \cdots, \Omega_N\}$, such that the ith unit is the winner for any input vector $\mathbf{x}$ in the region Ω_i. If we consider those vectors which are within the region Ω_i form a *cluster* and the state of the ith unit S_i as the *prototype* for these vectors, we see that a CL model is just a clustering algorithm [3], [4], in which N prototypes are generated for the set of vectors in Ω. Indeed, many existing CL models have been proposed to solve different forms of clustering problems, including pattern and speech recognition [2], [5]–[7], classification [8], [9], vector quantizer design [10]–[14] and probability density function (pdf) estimation [15], [16]. In all these CL models, the weight vector $\mathbf{w_i}$ is an important component (but may not be the only one) in specifying the state of the ith neuron S_i. At the tth learning step, a neuron is then rewarded by modifying its weight vector $\mathbf{w}(t)$ according to the Grossberg's learning rule [17], i.e.,

$$\triangle\mathbf{w}(t) = \alpha(t)\gamma_i(\mathbf{x}(t), S(t), t)(\mathbf{x}(t) - \mathbf{w}(t)) \tag{1}$$

where $\alpha(t)$ is a nonnegative learning rate, and $\gamma_i(\mathbf{x}(t), S(t), t)$ is the competition score of the ith unit mentioned earlier. For the convergence of weight vectors, it is very often required that $\alpha(t)$ tends to zero as $t \to \infty$. Note that the larger the learning rate, the larger will be the reward. In fact, in the context of clustering, the weight vectors in these CL models are the prototypes being generated or learned for their respective clusters.

The simple competitive learning (SCL) [18] is probably the simplest CL model. The state of the ith unit is uniquely determined by its weight vector, i.e., $S_i = \{\mathbf{w_i}\}\ \forall i \in \{1, \cdots, N\}$. The winner k is the neuron satisfying

$$k = \arg \min_{i \in \{1, \cdots, N\}} \{|\mathbf{x} - \mathbf{w_i}|\} \tag{2}$$

where $\mathbf{x}$ is the input vector, $|\mathbf{y}|$ denotes the Euclidean norm of a vector $\mathbf{y}$, and ties are broken arbitrarily. Then, only the winner's weight vector $\mathbf{w_k}$ is modified by the Grossberg's learning rule in (1) with a positive learning rate. Despite its simplicity, it was noticed by Rumelhart and Zipser [6] that when input vectors were not drawn from a simply connected region or when weight vectors were not initially located uniformly within the aforementioned region, some of the neurons in the SCL will never win. We refer to this as the *neuron underutilization* problem in this paper.

The objective of this paper is to suggest a class of CL models called the multiplicatively biased competitive learning

(MBCL) model and to prove that it can avoid neuron underutilization under appropriate conditions. We organize this paper as follows. In the next section, we will define our notion of neuron underutilization, and discuss the implications of neuron underutilization, namely, suboptimal performance and sensitivity to initial conditions. CL models which aim at avoiding this problem, either explicitly or implicitly, will be reviewed. We will discuss two reasons for studying the MBCL. First, it can be efficiently implemented in a sequential environment, which is crucial for practical applications of neural networks. Actually, all existing instances of the MBCL have complexities close to that of the SCL. Second, it can avoid neuron underutilization without any complicated control of its parameters during operations. In Section III, we will propose a theorem stating the sufficient conditions for an MBCL to avoid neuron underutilization with probability one, and present the proof to support our claim. Finally, we will conclude this paper with some discussions. It is our understanding that we are the first to explicitly prove that a class of CL models can avoid the neuron underutilization problem. By introducing this theorem, we hope that new instances of MBCL can be developed which can avoid neuron underutilization problem and yet suitable for specific applications.

II. Neuron Underutilization Problem

A. Definition

Let us formally state the definition of neuron underutilization we have adopted.

Definition 1: Given a CL model such that it partitions an input space Ω into regions $\{\Omega_1(t), \cdots, \Omega_N(t)\}$ at any time t. Let us denote an indicator function $I_i(t)$ such that

$$I_i(t) = \begin{cases} 1 & \text{if } \mathbf{x}(t) \in \Omega_i(t) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then, when the following is true, the CL model does not have neuron underutilization problem:

$$\forall i,\ 0 < t' < \infty,\ \exists \text{ finite integer } T \qquad \text{s.t. } \sum_{t=t'}^{t'+T} I_i(t) > 0.$$

□

Equivalently, we have

$$\forall i,\ \lim_{T\to\infty} \sum_{t}^{T} I_i(t) = \infty. \tag{4}$$

Conversely, when there is a neuron j, such that

$$\exists 0 < t' < \infty, \qquad \text{s.t. } \forall t > t',\ I_j(t) = 0 \tag{5}$$

then, we say that the neuron is *underutilized*.

In other words, a neuron is said to be underutilized when it is not *persistently excited*, i.e., it wins (or being excited) for finitely many times given infinitely many learning times. This idea is borrowed from the idea of persistency of excitation in parameter estimation literature, see for example [19], and has appeared in the work of Kosmatopoulos and Christodoulou [20] as a sufficient condition in their proof on the convergence of a class of algorithms called the learning vector quantization (LVQ) algorithms. Their LVQ algorithm can be considered as a CL model in our sense, in which $\gamma_i(\mathbf{x}, S, t)$ is either 0 or 1.

Let us illustrate the physical meaning of the above definition by considering the SCL. Suppose the input vectors are sampled independently from the input space $\Omega = \Re^K$ (i.e., the K-dimensional Euclidean space) with a stationary probability density function $p(\mathbf{x})$. Furthermore, for illustration purpose, let us suppose that the learning rate $\alpha(t)$ is zero for $t > T$, where T is finite. Then, for $t > T$, it must be the case that $\triangle \mathbf{w_i}(t) = 0$ (i.e., it is forced to "converge"). Since weight vectors are unchanged for all $t > T$, we see that the SCL gives a stable partition of the input space $\Re^K$ and is denoted by $\{\Omega_1, \cdots, \Omega_N\}$. If there is a neuron i such that $\mathcal{P}(\Omega_i) = 0$, where $\mathcal{P}(Y)$ denotes the probability of the input $\mathbf{x}$ being sampled from the set Y i.e.,

$$\mathcal{P}(Y) = \int_Y p(\mathbf{x})\, d\mathbf{x} \tag{6}$$

neuron i cannot win infinitely often in time (since T is finite). Hence, according to the definition, neuron i is underutilized. On the other hand, if $\mathcal{P}(\Omega_i) \neq 0\ \forall i$, any neuron can win infinitely often in time and hence no neuron underutilization problem exists in this case.

Conventionally, a neuron is considered to be underutilized when it never wins. Note that our definition is more general than conventional ones, and we expect it is applicable to general situations. For example, when the input vectors are sampled from a slowly time varying probability density function $p(\mathbf{x}, t)$, then it may be possible that a neuron i wins for only a finite number of times at the early stage of learning since the pdf has "shifted" out of its respective region Ω_i. According to our definition, neuron i is underutilized, but if we use the conventional definition, it is not. However, if one aims at tracking a slowly time-varying pdf using a CL model (e.g., pdf estimation on a slowly time-varying pdf), the fact that neuron i is underutilized means that suboptimal performance will be resulted some time in the future.

B. Implications of Neuron Underutilization

Conceptually, when there is no underutilized neuron, it means that the CL model makes the maximal usage of the resources (neurons) being provided. There are two consequences of neuron underutilization, namely suboptimal performance and sensitivity to initial conditions.

Neuron underutilization will lead to suboptimal performance in some applications, and has been noticed by many researchers in applications including classification [8], [9], [21], vector quantizer design [10], [11] and pdf estimation [15], [16]. Let us elaborate the effect of neuron underutilization problem on each of these applications as follows. In pdf estimation, input vectors are assumed to be sampled from a stationary distribution $p(\mathbf{x})$, and the objective is to assign weight vectors (prototypes or representative vectors) such that their distributions approximate as close to $p(\mathbf{x})$ as possible [i.e., to form a nonparametric model of $p(\mathbf{x})$]. Consequently, when it is given a sufficiently large number of neurons, regions in the vector space with zero probability must be assigned with

no weight vector. However, when a CL model like the SCL is applied to pdf estimation, if a neuron k is underutilized according to (5), the neuron does not win after some time instance t'. Consequently, $\mathcal{P}(\Omega_k(t)) = 0$ for $t > t'$, and the weight vector $\mathbf{w_k}$ is situated at the region of zero probability. Hence, at least one neuron is effectively useless in this pdf estimation application, which corresponds to suboptimal performance. Likewise, in data clustering for classification purpose, if one is targeted at partitioning the set of vectors into a predetermined number of clusters as in [8], [9], and [21], an underutilized neuron means that there is a region without any data vector. This means that the effective number of clusters is reduced. The codebook design problem in vector quantizer design [22] is just a data clustering problem where the number of partitions on a set of training vectors is the codebook size. Each weight vector is a codebook, and is used to represent the set of training vectors in its respective Voronoi region. If approaches susceptible to neuron underutilization problem like the SCL is employed, it is possible that at least one weight vector does not represent any training vector, which means that the overall distortion introduced must be increased. Hence, in these applications, a necessary condition for optimal performance is to avoid the neuron underutilization problem.

In addition, if a neuron in a CL model is underutilized, its performance may be very sensitive to initial conditions, as in the case of the SCL. Indeed, in an unsupervised learning model, since there is little prior knowledge concerning the distribution of input vectors to be learned, the final state of the unsupervised learning model should not, ideally, be affected by initial conditions, provided that it is learned for sufficiently long time. Hence, insensitivity to initial conditions seems to be a reasonable prerequisite for an unsupervised CL model. In existing CL models, whenever a neuron wins, its weight vector must be updated using (1) with nonnegative learning rate. Referring to the convergence analysis of (1) by Clark and Ravishankar [23], they proved that when input vector $\mathbf{x}$ is sampled from a set of training patterns Y independent of previously sampled vectors, the weight vector $\mathbf{w}$ will converge to the probabilistic centroid of Y in probability[2] if and only if $\lim_{t\to\infty} \alpha(t) = 0$ and the series $\lim_{T\to\infty} \sum_{t=0}^{T} \alpha(t)$ diverges. Their work implies that the converged weight vector will be independent of its initial state since the probabilistic centroid of Y can be arbitrary, i.e., insensitivity to initial conditions. Since learning rate must be finite to be physically realizable, the above conditions mean that if the weight vector has to converge to the probabilistic centroid, it is necessary for the weight vector to be updated infinitely often in time (i.o.t. for short).

The work of Kosmatopoulos and Christodoulou [20] extends that of Clark and Ravishankar [23] to the general situation in their LVQ algorithms, in which there are more than one neuron. In their LVQ algorithms, weight vector of a neuron is either updated or unchanged [i.e., $\gamma_i(\mathbf{x}, S, t)$ is either zero or one] by the Grossberg's learning rule in (1) at each learning step. Note that the SCL is an instance of the LVQ algorithm. They proved, under the assumption that any neuron is persistently excited (i.e., according to Definition 1) and other assumptions similar to those in the work of Clark and Ravishankar [23], that their LVQ algorithm will converge with probability one. However, as they have pointed out in their work, the assumption on persistency of excitation of any neuron may not be satisfiable, e.g., in the SCL in which neuron underutilization is possible. Note that the MBCL we will define in Section III satisfies the requirement of their LVQ algorithm. Hence, by proving that an MBCL satisfies Definition 1, we can apply their theorem to prove that the MBCL will converge with probability one.

However, we must point out that in some applications, avoiding neuron underutilization problem may not be necessary. For example, the SOM from Kohonen [2] can be considered a CL model performing clustering with constraints on weight vectors. Weight vectors are constrained to a fixed topology, and when the underlying pdf $p(\mathbf{x})$ does not match this topology, some neurons may never win [24]. Since the objective of the SOM is to preserve topological relationship, this may conflict with the requirement to avoid neuron underutilization.

C. Approaches in Avoiding Neuron Underutilization

In recent years, many CL models were proposed to avoid implicitly or explicitly the neuron underutilization problem. Basically, they can be considered as variants of two models from Grossberg [18]: the partial-contrast (PC) model and the variable-threshold (VT) model.

1) Variants of the PC Model: In the PC model [18], each neuron has a weight vector only. Weight vectors of losers are updated [using (1)] as well as that of the winner, although with a smaller learning rate. Suppose that the pdf is a stationary one, the concept is that by updating losers as well as winner, weight vectors of losers may be dragged toward the convex hull of the pdf. However, if there are portions inside the convex hull of the pdf with zero probability, neuron underutilization problem may still exist.

The self-organizing map (SOM) from Kohonen [2] can be considered as a variant of the PC model, in which the learning rate of a neuron is a function of the "distance" between the neuron and the winner, where "distance" is defined according to a predefined topological relationship among neurons. The learning rate is a decreasing function of the "distance" from the winner. As we have previously discussed, the SOM with fixed topology may not avoid neuron underutilization. Accordingly, Kangas *et al.* [24] suggested an approach to dynamically change the topological relationship among neurons, so as to give a better approximation to the underlying pdf.

In the work of Pal *et al.*, they considered a generalization of the SCL called the generalized learning vector quantization (GLVQ) [8] (the LVQ they were referring in their paper is just the SCL in our paper; however, we prefer the name SCL since the LVQ may lead to confusion with the supervised learning algorithm from Kohonen [1]) to avoid neuron underutilization

[2] i.e., suppose $\mathbf{z}$ is the probabilistic centroid of Y, then for every $\varepsilon > 0$

$$\lim_{t\to\infty} \mathcal{P}(\|\mathbf{w}(t) - \mathbf{z}\| < \varepsilon) = 1 \tag{7}$$

where $\|\mathbf{x}\|$ is a norm for a vector $\mathbf{x}$.

problem, and demonstrated its performance in clustering for classification. In the GLVQ, learning rates are functions of distances between the input vector $\mathbf{x}$ and all weight vectors. Although the GLVQ minimizes a well-defined energy function, it was later demonstrated by Gonzalez *et al.* [21] that the GLVQ resembles the behavior of the SCL under some conditions, which means that neuron underutilization in the GLVQ is indeed possible.

In [9], Karayiannis *et al.* suggested a remedy to the GLVQ to avoid neuron underutilization by incorporating the concept of fuzziness [25], [26]. The concept of fuzziness has also been applied by Chung and Lee [27] to explicitly avoid the neuron underutilization problem. Other approaches like the partial-distortion-weighted fuzzy competitive learning algorithm [28] and the fuzzy learning vector quantization (FLVQ) in [29] have also incorporated the concept of fuzziness. In these approaches, the learning rate of each neuron is a function of distances between the input vector $\mathbf{x}$ and all weight vectors, such that the winner has the largest learning rate.

In [12], Martinetz *et al.* suggested the neural-gas network which can be considered as a variant of the PC model. The learning rate is a function of the rank of the distances of the input vector from all weight vectors, such that the smaller the distance, the larger will be the learning rate. They have proved that when N tends to infinity, and under suitable conditions including smoothness requirement on the pdf $p(\mathbf{x})$, the stationary distribution of weight vectors corresponds to the distribution of codewords in an asymptotically optimal vector quantizer, i.e., $\propto p(\mathbf{x})^{K/(K+2)}$ where K is the vector dimension [30]. Hence, they have implicitly shown that neuron underutilization problem does not occur in the neural-gas. However, this occurs when N tends to infinity.

2) The VT Model: In the VT model [18], the concept is to increase the relative fitness of a unit when it fails to win competition, while fitness decreases for frequent winning. According to existing literature, we can classify CL models which implement the idea of VT model into two types: the additively biased competitive learning (ABCL) model and the MBCL model. Indeed, in existing variants of the VT model, only the winner's weight vector is updated, as different from the PC model. We define these two models with reference to the general CL model introduced in Section I. In both of these models, the state of the ith neuron is determined by a scalar b_i, which we called the *bias factor* and a weight vector $\mathbf{w_i}$, i.e., $S_i = \{b_i, \mathbf{w_i}\}$. Furthermore, if k is the winner in competition, the competition score of the ith neuron is given by

$$\gamma_i(\mathbf{x}, S, t) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In the ABCL, the winner k at the tth learning step satisfies

$$k = \arg\min_i \{d(\mathbf{x}(t), \mathbf{w_i}(t)) + A_i(S(t), t)\} \tag{9}$$

where $d(\mathbf{x}, \mathbf{w})$ is the distance between the two vectors $\mathbf{x}$ and $\mathbf{w}$, $S(t)$ is the current state of the CL model, i.e., $S(t) = \{S_1(t), \cdots, S_N(t)\}$, and $A_i(S(t), t)$ is a scalar-valued function affecting which neuron wins. Note that the influence is in the form of addition. Weight vectors are updated according to

$$\begin{aligned} \forall i,\ \mathbf{w_i}(t+1) &= \mathbf{w_i}(t) + \triangle \mathbf{w_i}(t) \\ &= \mathbf{w_i}(t) - \alpha(t)\gamma_i(\mathbf{x}(t), S(t), t) \\ &\quad \cdot \frac{\partial d(\mathbf{x}(t), \mathbf{w_i}(t))}{\partial \mathbf{w_i}} \end{aligned} \tag{10}$$

while bias factors are updated according to

$$\forall i,\ b_i(t+1) = U_i(k, \mathbf{x}(t), S(t), t). \tag{11}$$

Note that only the winner's weight vector is modified.

In the MBCL, the winner k at the tth learning step satisfies

$$k = \arg\min_i \{M_i(S(t), t)\, d(\mathbf{x}(t), \mathbf{w_i}(t))\} \tag{12}$$

where $d(\mathbf{x}, \mathbf{w})$ and $S(t)$ have their usual meanings, and $M_i(S(t), t)$ is a scalar-valued function affecting which neuron wins. Note that the influence is in the form of multiplication. Same as the case of the ABCL, weight vectors are modified according to (10) and bias factors are modified according to (11).

The conscience competitive learning (CCL) suggested by DeSieno [15] to solve the neuron underutilization problem is an instance of the ABCL, such that

$$d(\mathbf{x}, \mathbf{w}) = |\mathbf{x} - \mathbf{w}|^2 \tag{13a}$$

$$A_i(S(t), t) = -\zeta\left(\frac{1}{N} - b_i(t)\right) \tag{13b}$$

$$U_i(k, \mathbf{x}(t), S(t), t) = b_i(t) + \beta(\gamma_i(\mathbf{x}(t), S(t), t) - b_i(t)) \tag{13c}$$

where β and ζ are constants.

However, it has been recently demonstrated by Chen and Chang [16] that the *conscience parameter* ζ in (13b) in the CCL must not be too large, or else competition according to (13b) becomes independent of the distance measure $|\mathbf{x} - \mathbf{w}|^2$. Neither can it be too small, nor else the bias may not have any effect. The former case gives rise to what they referred to as *neuron tangling phenomenon,* while the latter case gives rise to neuron underutilization problem. They investigated a number of cases, and stated that the value of ζ in (13b) that makes the "conscience" effective depends heavily on the size and the location of the input domain, the number of neurons, the initial locations of neurons, etc. Since such information, except the number of neurons, is not known *a priori,* they suggested the adaptive conscientious competitive learning (ACCL) to adaptively vary the value of ζ. Although they showed that the ACCL had better performance than the CCL, they did not prove explicitly that neuron underutilization problem could be avoided with certainty.

Two versions of the frequency sensitive competitive learning (FSCL) which are instances of the MBCL model were proposed by Ahalt *et al.* [10] (we called it FSCL1) and by Krishnamurthy *et al.* [11] (we called it FSCL2), and are characterized by the following equations:

$$d(\mathbf{x}, \mathbf{w}) = |\mathbf{x} - \mathbf{w}|^2 \tag{14a}$$

$$M_i(S(t), t) = f(b_i(t)) \tag{14b}$$

$$U_i(k, \mathbf{x}(t), S(t), t) = b_i(t) + \gamma_i(\mathbf{x}(t), S(t), t). \tag{14c}$$

In the FSCL1, $f(u) = u$ while $f(u) = u^{\beta \exp\{-t/T\}}$ where β and T are constants at the tth learning step.

Recently, there are two other instances of the MBCL which are based on the equidistortion principle [30]. The distortion equalized competitive learning (DECL) from Butler and Jiang [13] has the form

$$d(\mathbf{x}, \mathbf{w}) = |\mathbf{x} - \mathbf{w}|^2 \tag{15a}$$

$$M_i(S(t), t) = \frac{b_i(t)}{\sum_j b_j(t)} \tag{15b}$$

$$U_i(k, \mathbf{x}(t), S(t), t) = b_i(t) + \gamma_i(\mathbf{x}(t), S(t), t) \cdot \frac{b_k(t)}{\sum_j b_j(t)}\, d(\mathbf{x}(t), \mathbf{w_k}(t)) \tag{15c}$$

$$b_i(0) = 1. \tag{15d}$$

The distortion sensitive competitive learning (DSCL) from Choy and Siu [14] is given by

$$d(\mathbf{x}, \mathbf{w}) = |\mathbf{x} - \mathbf{w}|^r \text{ where } r \text{ is a positive constant} \tag{16a}$$

$$M_i(S(t), t) = b_i(t) \tag{16b}$$

$$U_i(k, \mathbf{x}(t), S(t), t) = b_i(t) + \gamma_i(\mathbf{x}(t), S(t), t) \cdot d(\mathbf{x}(t), \mathbf{w_k}(t)) \tag{16c}$$

$$b_i(0) > 0. \tag{16d}$$

According to existing ABCL's and MBCL's, only the winner's weight vector is updated, which means that large computational savings are possible in their sequential realizations as compared to variants of PC model. Suppose there are N neurons and each weight vector has K dimensions. The complexity in determining the winner in an ABCL or an MBCL is comparable to or even lower than that in calculating learning rates in existing variants of the PC model. In fact, a short cut method exists [14] which can reduce substantially the number of operations in determining the winner in existing MBCL's. In addition, each learning step updates $N+K$ scalars at most (in the CCL and ACCL), while only $N+1$ scalars at its minimum (in the FSCL, DECL, and DSCL) in existing variants of the VT model. As compared to variants of PC model, at most NK scalars have to be modified since learning rates on each neuron can be nonzero. Hence, large computational savings are possible.

Although the ABCL can be as computationally efficient as that of the MBCL, we prefer the latter for the sake of simplicity. In the ACCL from Chen and Chang [16], the adaptive rule which adjusts the conscience parameter ζ in (13b) has introduced extra parameters into their approach which must be specified. On the other hand, we will prove that approaches as simple as the FSCL1 [10] and the DSCL [14] can avoid neuron underutilization problem.

III. Theorem and Proof

A. Preliminaries

Let us define the MBCL under consideration in this paper explicitly. The MBCL has N neurons, which is finite. The state of the ith neuron at the tth learning step (where $t \in \{0, 1, \cdots\}$) is denoted by $S_i(t) = \{b_i(t), \mathbf{w_i}(t)\}\ \forall i \in \{1, \cdots, N\}$, where $b_i(t) \in \Re$ is its bias factor and $\mathbf{w_i}(t) \in \Re^K$ is its weight vector. The state of the MBCL is then given by $S(t) = \{S_1(t), \cdots, S_N(t)\}$. For notational simplicity, S will be used to denote the current state of the MBCL. Similarly, for the ith neuron, S_i will denote its current state, b_i will denote its current bias factor and $\mathbf{w_i}$ will denote its current weight vector. The bias factor under the current state S will also be denoted as $b_i(S)$ or $b_i(S(t))$.

The index of the unique winner is given by

$$\eta(\mathbf{x}, S) = \arg\min_i \{M_i(S)\, d(\mathbf{x}, \mathbf{w_i})\} \tag{17}$$

where $\mathbf{x} \in \Re^K$ is the input vector when the MBCL is in state S. An indicator function is defined, such that

$$\gamma_i(\mathbf{x}, S) = \begin{cases} 1 & \text{if } i = \eta(\mathbf{x}, S) \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The distance measure is given by $d(\mathbf{x}, \mathbf{w}) = G(\mathbf{x} - \mathbf{w})$, where $G(\cdot)$ is a scalar-valued function satisfying

$$G(\mathbf{u}) \le G(\mathbf{v}) \Leftrightarrow \|\mathbf{u}\| \le \|\mathbf{v}\| \qquad \forall\, \mathbf{u}, \mathbf{v} \in \Re^K \tag{19a}$$

$$G(a \cdot \mathbf{u}) = g(a) G(\mathbf{u}) \qquad \forall\, a \ge 0,\ \mathbf{u} \in \Re^K \tag{19b}$$

$g(\cdot)$ is a strictly increasing scalar-valued function such that $g(0) = 0$ and $g(1) = 1$, and $\|\cdot\|$ is a norm in $\Re^K$. [For example, $G(\mathbf{u}) = \|u\|^r$ where $r > 0$ is a possible distance.] Only the state of the winner will be updated, that is,

$$\triangle \mathbf{w_i} = -\alpha(t) \gamma_i(\mathbf{x}, S)\, \frac{\partial\, d(\mathbf{x}, \mathbf{w_i})}{\partial \mathbf{w_i}} \tag{20a}$$

$$\triangle b_i = \gamma_i(\mathbf{x}, S) F_i(\mathbf{x}, S) \tag{20b}$$

where $F_i(\mathbf{x}, S)$ denotes the amount of increase of the bias factor, b_i, when neuron i wins.

At each state S, the input space $\Re^K$ is partitioned exhaustively into N disjoint regions, $\{\Omega_1(S), \cdots, \Omega_N(S)\}$, where

$$\Omega_i(S) = \{\mathbf{x} \in \Re^K \colon M_i(S)\, d(\mathbf{x}, \mathbf{w_i}) \le M_j(S)\, d(\mathbf{x}, \mathbf{w_j})\ \forall i \ne j\}\ \forall i \in \{1, \cdots, N\}. \tag{21}$$

Note that any point on the boundaries between regions is assigned to a unique arbitrary region. We will also use $\Omega_i(t)$ or Ω_i to denote the region corresponding to the ith neuron at the tth learning step.

Each region $\Omega_i(S)$ can be constructed by the intersection of finitely many sets, that is,

$$\Omega_i(S) = \bigcap_{j \ne i} \Gamma_{ij}(S),$$

where

$$\Gamma_{ij}(S) = \{\mathbf{x} \in \Re^K \colon M_i(S)\, d(\mathbf{x}, \mathbf{w_i}) \le M_j(S)\, d(\mathbf{x}, \mathbf{w_j})\}. \tag{22}$$

Let us define the following notations for the convenience of our discussion. Let A and B denote two sets, then $A \backslash B$ denotes the set of elements in A but not in B. For a set $A \subseteq \Re^K$, its volume is denoted by $V(A)$. Given a countable set A, the number of elements will be denoted as $|A|$. The probability of an event B in a probability space will be

denoted by $\mathcal{P}(B)$, while the conditional probability of this event given m random variables $\mathbf{y_1}, \cdots, \mathbf{y_m}$ is denoted by $\mathcal{P}(B|\mathbf{y_1}, \cdots, \mathbf{y_m})$. For a sequence with values $\{a_1, a_2, \cdots\}$ such that $\forall i > j$, $a_i \geq a_j$, and $\lim_{i\to\infty} a_i = a$, then we denote this by the notation $a_i \uparrow_{i\to\infty} a$. Similarly, the notation $a_i \downarrow_{i\to\infty} a$ means that $\forall i > j$, $a_i \leq a_j$, and $\lim_{i\to\infty} a_i = a$.

For our subsequent analysis, let us introduce the following definitions.

Definition 2: Given a function $f(\mathbf{x})$: $\Re^K \to \Re$, we define its *support*, $\Upsilon_{f(\cdot)}$, which is the smallest set such that

$$\int_{\Re^K \setminus \Upsilon_{f(\cdot)}} f(\mathbf{y})\, d\mathbf{y} = 0. \tag{23}$$

□

Definition 3: Let A be a set. Its *connected decomposition* is a partition denoted as $D_A = \{A_1, \cdots, A_m\}$ where m is the number of elements, such that

1) $\forall i$, A_i is connected, and
2) $\forall i \neq j$, $A_i \cup A_j$ is not connected.

□

Definition 4: We define $\Pi_S = \{S: \forall i, b_i(S) > 0\}$, which is the set of permissible states of the MBCL satisfying conditions in Theorem 1. □

Definition 5: The sequence $\{x_n\}$ converges to x almost surely (a.s.) or with probability one (w.p. 1) if

$$\mathcal{P}\Big(\lim_{n\to\infty} x_n = x\Big) = 1. \tag{24}$$

□

B. The Theorem

Theorem 1: Suppose that the following conditions are satisfied for the MBCL defined above.

1) The initial state $S(0)$ of the MBCL is bounded, i.e., all weight vectors and bias factors are bounded, and $\forall i$, $b_i(0) > 0$.
2) The learning rate, $\alpha(t)$, is nonnegative for all time $t \geq 0$, and is sufficiently small that
$$d(\mathbf{x}, \mathbf{w_i} + \triangle\mathbf{w_i}) \leq d(\mathbf{x}, \mathbf{w_i}) \tag{25}$$
where $\triangle\mathbf{w_i}$ is given by (20a).
3) Let $\{\mathbf{x}(0), \mathbf{x}(1), \cdots\}$ be a sequence of independent identically distributed random vectors, whose probability density function is given by $p(\mathbf{x})$, such that
   a) $p(\mathbf{x})$ is Lipschitz continuous, i.e.,
   $$\forall \mathbf{x_1}, \mathbf{x_2}, |p(\mathbf{x_1}) - p(\mathbf{x_2})| \leq h\|\mathbf{x_1} - \mathbf{x_2}\| \tag{26}$$
   where $h \geq 0$ is the Lipschitz constant, and $\|\cdot\|$ is a norm in $\Re^K$.
   b) $\Upsilon_{p(\cdot)}$ is bounded, such that there exists a hypersphere $B(\mathbf{c}, C_R) = \{\mathbf{x}: d(\mathbf{x}, \mathbf{c}) < C_R\}$ with $\Upsilon_{p(\cdot)} \subseteq B(\mathbf{c}, C_R)$.
4) For any state $S \in \Pi_S$:
   a) there exists positive finite C_F such that
   $$\forall \mathbf{x} \in B(\mathbf{c}, C_R),\ C_F \geq F_i(\mathbf{x}, S) \geq 0, \quad \text{and} \tag{27}$$
   b) either there is no solution to the equation $F_i(\mathbf{x}, S) = 0$ [i.e., $\Upsilon^{\mathbf{c}}_{F_i(\cdot, S)} = \emptyset$], or its solutions correspond to finitely many isolated singularities [i.e., $D_{\Upsilon^{\mathbf{c}}_{F_i(\cdot, S)}} = \{\{\mathbf{y_1}\}, \cdots, \{\mathbf{y_m}\}\}$ where m is finite].
5) For any state $S \in \Pi_S$:
   a) $M_i(S) > 0 \forall i$, and
   b) $\forall i, j$, $M_i(S)/M_j(S) = Q_{ij}(b_i/b_j)$, where $Q_{ij}(r)$ is continuous and strictly increasing in r, such that $\lim_{r\to 0} Q_{ij}(r) = 0$. [Note that since $Q_{ij}(r) = 1/(Q_{ji}(1/r))$, it must be the case that $\lim_{r\to\infty} Q_{ij}(r) = \infty$.]

Then, as $t \to \infty$, the MBCL will not have any underutilized neuron w.p. 1. □

C. The Proof

Before presenting the detailed mathematical proof, let us loosely describe our idea in proving Theorem 1 in the following. Suppose the learning rates for all weight vectors are zero, such that $\mathbf{w_i}$ is unchanged during learning. Furthermore, for simplicity, suppose that there is exactly one neuron i being underutilized, while all the other neurons are excited persistently. Note that according to Condition 4) in Theorem 1, whenever a neuron wins, its bias factor will be increased by a quantity which must be nonzero with probability one. Hence, after sufficiently long time, the bias factor of this neuron b_i will be arbitrarily smaller than bias factor of any other neuron, and that $b_i(t)/b_j(t)$ will be arbitrarily small for sufficiently large t for any $i \neq j$. Note that this ratio is nonnegative. Hence if this situation persists indefinitely, this ratio will tend to zero.

In order to understand the consequence of this ratio tending to zero, we have to analyze the relationship between the ratio $M_i(S)/M_j(S)$ $\forall j \neq i$ and the volume of the region $\Omega_i(t)$. Note that the region $\Omega_i(t)$ is the intersection of $N-1$ sets $\Gamma_{ij}(t)$ as defined in (22). Since the weight vector is not changed during learning, a decrease in the volume of any set $\Gamma_{ij}(t)$ will lead to a reduction in the volume of the region $\Omega_i(t)$. Consider the boundary of the set $\Gamma_{ij}(t)$, which is denoted as

$$\partial\Gamma_{ij}(t) = \big\{\mathbf{x} \in \Re^K: M_i(S)\, d(\mathbf{x}, \mathbf{w_i}) = M_j(S)\, d(\mathbf{x}, \mathbf{w_j})\big\}. \tag{28}$$

We refer to this as the *decision surface* of $\Gamma_{ij}(t)$. The shape of this surface is related to the ratio $M_i(S)/M_j(S)$—concave relative to the weight vector $\mathbf{w_i}$ when this ratio is larger then one, while convex relative to the weight vector $\mathbf{w_i}$ when this ratio is smaller than one. To visualize the effect of this ratio, we consider two weight vectors $\mathbf{w_1}$ and $\mathbf{w_2}$ in the two-dimensional case, and their decision surface $\partial\Gamma_{12}(t)$. Fig. 1 shows the decision surfaces for different ratios of $M_1(S)/M_2(S)$. It is obvious that when this ratio is greater than 1, a larger ratio means a more convex (relative to $\mathbf{w_1}$) surface, and hence, $\Gamma_{12}(t)$ decreases in volume. On the other hand, when this ratio is smaller than one, a smaller ratio means a more concave surface, and hence, $\Gamma_{12}(t)$ increases in volume. Extending this idea to higher dimensions, we see that the volume of $\Gamma_{ij}(t)$ has a tendency to increase when the ratio $M_i(S)/M_j(S)$ decreases.

According to the requirement in Condition 5) in Theorem 1, this ratio decreases when the ratio $b_i(t)/b_j(t)$ decreases.

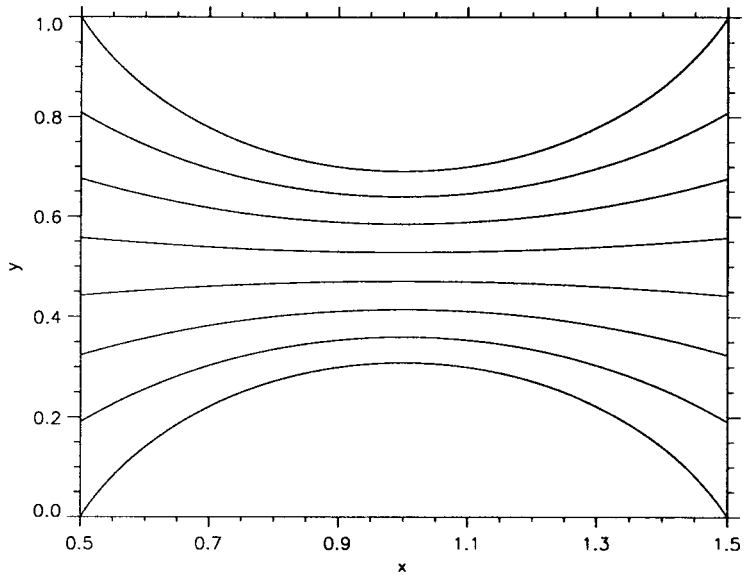

Fig. 1. Decision surface between two weight vectors $\mathbf{w_1} = \begin{bmatrix}1\\0\end{bmatrix}$ and $\mathbf{w_2} = \begin{bmatrix}1\\1\end{bmatrix}$ is denoted as $\partial\Gamma = \{\mathbf{x} \in \Re^2 : m_1|\mathbf{x} - \mathbf{w_1}| = m_2|\mathbf{x} - \mathbf{w_2}|\}$. The above are plots of decision surfaces with different ratio m_1/m_2: (from the bottommost curve to the topmost curve), respectively, 5, 3.1569, 1.9932, 1.2585, 0.7946, 0.5017, 0.3168, and 0.2. The horizontal axis is the x-dimension, while the vertical is the y-axis.

This means that after sufficiently long time, the volume of the region $\Omega_i(t)$ will increase to an arbitrarily large value. Since the pdf $p(\mathbf{x})$ is finite and the MBCL is initialized with finite state, it will only take finite amount of time for $\Omega_i(t)$ to "cut" the support of $p(\mathbf{x})$, such that $\mathcal{P}(\Omega_i(t')) > 0$ for some time t'. This mechanism makes neuron i wins again. Conceptually, this mechanism goes on indefinitely such that whenever a neuron is underutilized for sufficiently long time, it will win again.

Our proof requires the following lemma, which will be proved in Appendix A.

Lemma 1: Suppose an MBCL satisfies the conditions in Theorem 1. Then for any state $S \in \Pi_S$, any region $\Omega_i(S)$ satisfies $|D_{\Omega_i(S)}| < \infty$. □

Proof of Theorem 1: Since we require that $\forall\, i,\ b_i(0) > 0$ and that $\forall\, i,\ t,\ \triangle b_i(t) \geq 0$ (Condition 4a), it is obvious that a possible state, S, in the MBCL satisfies $\forall\, i,\ t,\ b_i(t) > 0$. This justifies the fact that Π_S is the set of permissible states of the MBCL under consideration, as defined in Definition 4.

Furthermore, according to Lemma 1, for the distance measure under consideration and for all possible states, each region is composed of finitely many connected subsets. This implies that the input space is also partitioned into finitely many connected subsets. Hence, if $F_i(\cdot)$ satisfies Condition 4), it is possible for the integral $\int_{\Omega_i} F(\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$ to be nonzero.

Outline of the Proof: According to (20b), we have

$$\begin{aligned}\lim_{t\to\infty} b_i(t) &= \lim_{T\to\infty} \sum_{t=0}^{T} \gamma_i(\mathbf{x}(t),\, S(t))F_i(\mathbf{x}(t),\, S(t)) + b_i(0)\\ &\leq C_F \lim_{T\to\infty} \sum_{t=0}^{T} \gamma_i(\mathbf{x}(t),\, S(t)) + b_i(0)\\ &\qquad \text{according to Condition 4.} \end{aligned} \tag{29}$$

Note that in case of the MBCL, $I_i(t)$ as defined in Definition 1 is equivalent to $\gamma_i(\mathbf{x}(t),\, S(t))$ as defined in (18). Since C_F and $b_i(0)$ are finite, we have

$$\lim_{t\to\infty} b_i(t) = \infty \Rightarrow \lim_{T\to\infty} \sum_{t=0}^{T} I_i(t) = \infty \tag{30}$$

which means that the MBCL has no underutilization problem. Hence, in order to prove the MBCL has no underutilization problem with probability one, it is sufficient to prove that

$$\forall\, i,\ \mathcal{P}\Big(\lim_{t\to\infty} b_i(t) = \infty\Big) = 1. \tag{31}$$

To prove (31), we will first prove that

$$\mathcal{P}\left(\lim_{t\to\infty} \sum_{i=1}^{N} b_i(t) = \infty\right) = 1. \tag{32}$$

Then, based on (32), we will prove (31).

Proof of (32): Let us consider the sum $\triangle b(t) = \sum_{i=1}^{N} \triangle b_i(t)$, which is a function of $\mathbf{x}(t)$ and $S(t)$ only. Indeed, given the same initial condition $S(0)$, $S(t)$ is uniquely determined by $\{\mathbf{x}(0), \cdots, \mathbf{x}(t-1)\}$. Hence, $\triangle b(t)$ is actually a function of $\{\mathbf{x}(0), \cdots, \mathbf{x}(t)\}$. Then, we have the following relationships.

1) According to Condition 4), we have $\triangle b(t) \geq 0$.
2) We show that $\mathcal{P}(\triangle b(t) = 0|S(t)) = 0,\ \forall\, t$ as follows.
   Since $\triangle b(t) = \sum_{i=1}^{N} \gamma_i(\mathbf{x}(t),\, S(t),\, t)F_i(\mathbf{x}(t),\, S(t))$, and that $\sum_{i=1}^{N} \gamma_i(\mathbf{x}(t),\, S(t),\, t) = 1$ for all t, $\mathbf{x}(t)$, and $S(t)$, we have

$$\begin{aligned}&\mathcal{P}(\triangle b(t) = 0|S(t))\\ &\quad = \sum_{i=1}^{N} \mathcal{P}(\mathbf{x} \in \Omega_i(S(t))|S(t))\\ &\qquad \cdot \mathcal{P}(F_i(\mathbf{x},\, S(t)) = 0|\mathbf{x} \in \Omega_i(S(t)))\\ &\quad = \sum_{i=1}^{N} \int_{\mathbf{x}\in\Upsilon^{\mathbf{c}}_{F_i(\cdot,\, S(t))}\cap\Omega_i(S(t))} p(\mathbf{x})\,d\mathbf{x}. \end{aligned} \tag{33}$$

According to Condition 4b), we have to consider two cases. When $\Upsilon^{\mathbf{c}}_{F_i(\cdot, S)} = \emptyset$, it is obvious that $\mathcal{P}(\triangle b(t) = 0 | S(t)) = 0$. In the other case, $\Upsilon^{\mathbf{c}}_{F_i(\cdot, S(t))}$ composes of finitely many isolated singularities. Due to the Lipschitz continuity requirement on $p(\mathbf{x})$ [Condition 3a)], we have

$$\forall \mathbf{y},\ \mathcal{P}(\{\mathbf{y}\}) = \lim_{r \to 0} \int_{\{d(\mathbf{x}, \mathbf{y}) < r\}} p(\mathbf{x})\, d\mathbf{x} = 0. \tag{34}$$

Hence, we have $\mathcal{P}(\triangle b(t) = 0 | S(t)) = 0$.

Assume that $\lim_{T \to \infty} \sum_{t=0}^{T} \triangle b(t) < \infty$, i.e., the sum converges. Since $\triangle b(t) \geq 0$, this implies that $\lim_{t \to \infty} \triangle b(t) = 0$. However, it has been shown above that $\mathcal{P}(\triangle b(t) = 0 | S(t)) = 0$ for all state $S(t)$ and time t. Hence, by contradiction, we conclude that (32) is true.

Proof of (31): Since (32) is valid, it is equivalent to say that we have the event E in the sample space of all possible sequences of input vectors $(\mathbf{x}(0), \mathbf{x}(1), \cdots)$ such that $\mathcal{P}(E) = 1$, where

$$E = \left\{ (\mathbf{x}(0), \mathbf{x}(1), \cdots) \colon \lim_{t \to \infty} \sum_{i=1}^{N} b_i(t) = \infty \right\}. \tag{35}$$

It is always possible to decompose this event into

$$E = \bigcup_{\substack{W \subseteq \{1, \cdots, N\} \\ W \neq \emptyset}} E_W \tag{36a}$$

where

$$E_W = \left\{ (\mathbf{x}(0), \mathbf{x}(1), \cdots) \colon \begin{array}{l} \lim_{t \to \infty} b_k(t) = \infty \\ \quad \forall k \in W, \text{ and} \\ \lim_{t \to \infty} b_i(t) < \infty \\ \quad \forall i \in \{1, \cdots, N\} \backslash W \end{array} \right\}. \tag{36b}$$

Note that $\mathcal{P}(E_\emptyset) = 0$ since (32) is valid. If it is true that $\mathcal{P}(E_{\{1, \cdots, N\}}) = 1$, the validity of (31) then follows. Hence, it is sufficient to prove that $\mathcal{P}(E_W) = 0, \forall W \subset \{1, \cdots, N\}$.

Let us consider the event E_W in (36b) in the following, where $W \subset \{1, \cdots, N\}$. We define the complement of W as $W^{\mathbf{c}} = \{1, \cdots, N\} \backslash W$. According to (36b) that $\lim_{t \to \infty} b_i(t) < \infty\ \forall i \in W^{\mathbf{c}}$, there exists a finite bound $b_{i(\max)}$ such that $\lim_{t \to \infty} b_i(t) \leq b_{i(\max)}\ \forall i \in W^{\mathbf{c}}$. In the following, where we have not explicitly defined, i is used as an index in the set $W^{\mathbf{c}}$ while k as an index in the set W.

For a neuron $k \in W$ at a particular time t, we have

$$\begin{aligned} &\mathbf{x} \in \Omega_k(S(t)) \cap B(\mathbf{c}, C_R) \\ &\quad \Rightarrow \forall i \in W^{\mathbf{c}},\ M_k(S(t))\, d(\mathbf{x}, \mathbf{w_k}(t)) \\ &\qquad < M_i(S(t))\, d(\mathbf{x}, \mathbf{w_i}(t)) \\ &\quad \Rightarrow \forall i \in W^{\mathbf{c}},\ d(\mathbf{x}, \mathbf{w_k}(t)) \\ &\qquad < Q_{ik}\left(\frac{b_i(t)}{b_k(t)} \right) d(\mathbf{x}, \mathbf{w_i}(t)). \end{aligned} \tag{37}$$

Let us derive an upper bound for $d(\mathbf{x}, \mathbf{w_k})\ \forall k \in W$, by deriving bounds for the two terms to the right of the above inequality.

The distance between any sampled vector $\mathbf{x}$ and weight vector $\mathbf{w}$ of any neuron [i.e., $d(\mathbf{x}, \mathbf{w})$] at time t is bounded by

$$C_D = \max \left\{ 2C_R, \max_j \sup_{\mathbf{x} \in B(\mathbf{c}, C_R)} d(\mathbf{x}, \mathbf{w_j}(0)) \right\}. \tag{38}$$

This bound is derived by considering two cases for any neuron j.

1) *The weight vector $\mathbf{w_j}$ is initialized within the hypersphere $B(\mathbf{c}, C_R)$.* In this case, it can either be the situation when neuron j wins at least once, or does not have a chance to win.
   a) In the former situation, its weight vector is updated according to (20a). Since the learning rate satisfies Condition 2), it must be the case that the distance between the weight vector $\mathbf{w_j}$ and the input vector $\mathbf{x}$ does not increase after learning. Since input vector $\mathbf{x}$ is in the hypersphere $B(\mathbf{c}, C_R)$ (Condition 3b), the weight vector after learning is still within the hypersphere.
   b) In the latter situation, its weight vector is never updated, and hence, it must be still inside the hypersphere for infinitely long time.

   Hence, the distance $d(\mathbf{x}, \mathbf{w})$ is bounded by the diameter of the hypersphere $B(\mathbf{c}, C_R)$, which corresponds to the first term in the bracket in (38).
2) *The weight vector $\mathbf{w_j}$ is initialized outside the hypersphere $B(\mathbf{c}, C_R)$.* Similar to the previous case, the neuron may win or may not win. Let us consider the extreme situation when the neuron never wins. Since its weight vector is never updated, the distance $d(\mathbf{x}, \mathbf{w})$ is bounded by $\sup_{\mathbf{x} \in B(\mathbf{c}, C_R)} d(\mathbf{x}, \mathbf{w}(0))$. This leads us to include the second term in the bracket in (38).

Since $\forall i \in W^{\mathbf{c}}$, $b_i(t) \leq b_{i(\max)}$ at any time t, we have $b_i(t)/b_k(t) \leq b_{i(\max)}/b_k(t)$. Owing to the increasing nature of $Q_{ik}(\cdot)$ [Condition 5)], we have

$$\begin{aligned} Q_{ik}\left(\frac{b_i(t)}{b_k(t)} \right) &\leq Q_{ik}\left(\frac{b_{i(\max)}}{b_k(t)} \right) \\ &\leq \max_{j \in W^{\mathbf{c}}, l \in W} Q_{jl}\left(\frac{b_{j(\max)}}{b_l(t)} \right) \\ &\leq \max_{j \in W^{\mathbf{c}}, l \in W} Q_{jl}\left(\max_{m \in W} \frac{b_{j(\max)}}{b_m(t)} \right). \end{aligned} \tag{39}$$

Substituting (38) and (39) into (37), we know that $\Omega_k(t) \cap B(\mathbf{c}, C_R)$ is bounded by $B(\mathbf{w_k}(t), \rho(S(t)))$, which is a hypersphere with radius[3] $\rho(S(t))$ centered at $\mathbf{w_k}(t)$, where

$$\rho(S(t)) = C_D \max_{j \in W^{\mathbf{c}}, l \in W} Q_{jl}(r_j(t))$$

with

$$r_j(t) = \max_{m \in W} \frac{b_{j(\max)}}{b_m(t)}. \tag{40}$$

[3] Note that distance is according to the measure $d(\mathbf{x}, \mathbf{y})$.

With respect to the two sets of neurons, W and $W^{\mathbf{c}}$, $\Upsilon_{p(\cdot)}$ is partitioned into two sets $\Omega_{\infty}(S(t))$ and $\Omega_{\overline{\infty}}(S(t))$ at a particular state $S(t)$, where

$$\Omega_{\infty}(S(t)) = \left\{ \bigcup_{k \in W} \Omega_k(S(t)) \right\} \cap \Upsilon_{p(\cdot)}$$
$$\Omega_{\overline{\infty}}(S(t)) = \left\{ \bigcup_{i \in W^{\mathbf{c}}} \Omega_i(S(t)) \right\} \cap \Upsilon_{p(\cdot)}. \tag{41}$$

Let us define the following:

$$\hat{\Omega}_{\infty}(S(t)) = \bigcup_{k \in W} B(\mathbf{w_k}(t), \rho(S(t)))$$
$$\hat{V}_{\infty}(t) = \sum_{k \in W} V(B(\mathbf{w_k}(t), \rho(S(t)))). \tag{42}$$

Then, we have the following upper bounds:

$$\hat{\Omega}_{\infty}(S(t)) \supseteq \Omega_{\infty}(S(t))$$

and

$$V(\Omega_{\infty}(S(t))) \leq V(\hat{\Omega}_{\infty}(S(t))) \leq \hat{V}_{\infty}(t). \tag{43}$$

Similarly, with the following definitions:

$$\check{\Omega}_{\overline{\infty}}(S(t)) = \Upsilon_{p(\cdot)} \backslash \hat{\Omega}_{\infty}(S(t))$$
$$\check{V}_{\overline{\infty}}(t) = V(\Upsilon_{p(\cdot)}) - \hat{V}_{\infty}(t) : \tag{44}$$

we have the following lower bounds:

$$\check{\Omega}_{\overline{\infty}}(S(t)) \subseteq \Omega_{\overline{\infty}}(S(t))$$

and

$$V(\Omega_{\overline{\infty}}(S(t))) \geq V(\check{\Omega}_{\overline{\infty}}(S(t))) \geq \check{V}_{\overline{\infty}}(t). \tag{45}$$

Since we are considering sequences of $\{\mathbf{x}(0), \mathbf{x}(1), \cdots\}$ from the event E_W in (36b), it is true that $\forall k \in W$, $\lim_{t \to \infty} b_k(t) = \infty$. Furthermore, $\triangle b_k(t)\ \forall t$ is bounded [i.e., $< C_F$ according to Condition 4a), such that $b_k(t)$ will not suddenly "jump" from finite to an infinite value], we have $r(t) \downarrow_{t \to \infty} 0$ in (40). From the increasing and continuous properties of $Q_{ik}(\cdot)$ [Condition 5)], we know that $\rho(S(t))$ in (40) decreases progressively with time. In fact, we have $\rho(S(t)) \downarrow_{t \to \infty} 0$. Hence, we have $\hat{V}_{\infty}(t) \downarrow_{t \to \infty} 0$ and $\check{V}_{\overline{\infty}}(t) \uparrow_{t \to \infty} V(\Upsilon_{p(\cdot)})$. It has to be recalled that $V(\Upsilon_{p(\cdot)}) \neq 0$ as implied by the Lipschitz continuous property of $p(\mathbf{x})$ [Condition 3a)].

Since $\triangle b_j(t) \geq 0\ \forall j$, in order for $\lim_{t \to \infty} b_i(t) < \infty$ (i.e., converging), it is necessarily true that $\lim_{t \to \infty} \triangle b_i(t) = 0$. Let us define $\triangle \bar{b}(t) = \sum_{i \in W^{\mathbf{c}}} \triangle b_i(t)$, and the event E'_W:

$$E'_W = \left\{ (\mathbf{x}(0), \mathbf{x}(1), \cdots) \in E_W \colon \lim_{t \to \infty} \triangle \bar{b}(t) = 0 \right\} \tag{46}$$

then we have $E_W \subseteq E'_W$ and $\mathcal{P}(E_W) \leq \mathcal{P}(E'_W)$.

There are two conditions when $\triangle \bar{b}(t) = 0$, namely, when either

1) $\mathbf{x}(t) \in \Omega_i(S(t))$ and $F_i(\mathbf{x}(t), S(t)) = 0$, where $i \in W^{\mathbf{c}}$; or
2) $\mathbf{x}(t) \in \Omega_k(S(t))$, where $k \in W$.

Hence, E'_W can be rewritten as

$$E'_W = E'_{W,1} \cup E'_{W,2}$$

where

$$E'_{W,1} = \Big\{ (\mathbf{x}(0), \mathbf{x}(1), \cdots) \in E_W \colon \mathbf{x}(t) \in \Upsilon^{\mathbf{c}}_{F_i(\cdot, S(t))} \cap\ \Omega_i(S(t))\ \forall i \in W^{\mathbf{c}} \text{ as } t \to \infty \Big\}$$
$$E'_{W,2} = \{ (\mathbf{x}(0), \mathbf{x}(1), \cdots) \in E_W \colon \mathbf{x}(t) \in \Omega_k(S(t)) \quad \forall k \in W \text{ as } t \to \infty \}$$
$$= \{ (\mathbf{x}(0), \mathbf{x}(1), \cdots) \in E_W \colon \mathbf{x}(t) \in \Omega_{\infty}(S(t)) \quad \text{as } t \to \infty \}, \tag{47}$$

with $\Omega_{\infty}(\cdot)$ as defined in (41).

Note that by applying the same argument in proving the integral in (33) evaluates to zero, we know that $\mathcal{P}(E'_{W,1}) = 0$ in (47). In addition, since $\hat{V}_{\infty}(t) \downarrow_{t \to \infty} 0$ and that $V(\Omega_{\infty}(S(t))) = V_{\infty}(t)$ is the upper-bounded by $\hat{V}_{\infty}(t)$ according to (43), it must be that $\lim_{t \to \infty} V(\Omega_{\infty}(S(t))) = 0$. Since $p(\mathbf{x})$ is Lipschitz continuous, we know that $\mathcal{P}(E'_{W,2}) = 0$ as well. Hence, we have $\mathcal{P}(E_W) = 0$.

This completes our proof. □

D. Discussions

As a consequence of no neuron underutilization according to Theorem 1, bias factors of all neurons of any MBCL satisfying the theorem will tend to infinity as time goes to infinity. This may pose problem in practical realization, especially when the number of learning steps t is very large. In order to alleviate this problem, we present an equivalent realization of an MBCL satisfying conditions in Theorem 1, using *normalized bias factors* c_i instead of bias factors b_i. Normalized bias factor c_i is defined as

$$\forall t > 0,\ i \in \{1, \cdots, N\},\ c_i(t) = \frac{b_i(t)}{t} \tag{48}$$

in which the corresponding bias factor is normalized by the number of learning steps, t. According to Condition 5) of Theorem 1, we have

$$\forall i, j,\ \frac{M_i(S)}{M_j(S)} = Q_{ij}\left(\frac{b_i}{b_j}\right) = Q_{ij}\left(\frac{c_i}{c_j}\right). \tag{49}$$

To realize the original MBCL using normalized bias factors, we redefine (17) such that the winner's index satisfies

$$\eta(\mathbf{x}, S) = \arg\min_i \left\{ Q_{i1}\left(\frac{c_i}{c_1}\right) d(\mathbf{x}, \mathbf{w_i}) \right\}. \tag{50}$$

(Since it is obvious that $b_i(t) > 0$ for all $t > 0$, $c_i(t)$ is nonzero for all $t > 0$. Hence, the choice of denominator is completely arbitrary. We choose neuron 1 as the denominator

for convenience.) Furthermore, from (20b), updating equation for the normalized bias factor, c_i, is given by

$$\triangle c_i(t) = -\frac{1}{1+t}(c_i(t) - \gamma_i(\mathbf{x}(t), S(t))F_i(\mathbf{x}(t), S(t))). \tag{51}$$

Weight vectors are updated in the same way as in the original MBCL (20a). It is obvious that this new realization gives exactly the same states as the original one, except that all bias factors are scaled down by t. Since these normalized bias factors are bounded for all $t > 0$ (as shown in Appendix B), we consider this to be a practical realization for prolonged learning. However, in each learning step, all normalized bias factors have to be updated, whereas in the original MBCL, only the bias factor of the winner needs to be updated. This increases learning time in sequential realization. Note that from our experimental results on the DSCL and FSCL1 [14], these MBCL's converge faster than a neural-gas network [12] and the DSCL performs better than all others. Indeed, their fast convergence property suggests that prolonged training may not be necessary in practice. Therefore, we expect that for practical usage in sequential environment, the original implementation of the MBCL is sufficient and more efficient.

According to Theorem 1, it holds even when the learning rate is zero. Although as we have loosely described in the previous section that the region $\Omega_i(t)$ of the ith neuron will eventually "cut" the support of the pdf $p(\mathbf{x})$ independent of the learning rate (as long as it is sufficiently small and nonnegative), the MBCL will be useless if weight vectors are not modified. On the other hand, since any neuron will eventually win, by using a learning rate that decreases at a sufficiently slow rate, it is always possible for the weight vector to move to the interior of the support of the pdf $p(\mathbf{x})$ from any initial conditions after sufficiently long time. Hence, this makes the MBCL insensitive to initial conditions as we have described in Section II.

Our approach to proving Theorem 1 demands for very loose smoothness requirement on the pdf $p(\mathbf{x})$. However, since we have to prove that the probability of the event $E'_{W,2}$ is zero in (47), "impulses" in the $p(\mathbf{x})$ are disallowed. Consider the case when $\lim_{t\to\infty} V(\Omega_\infty(S(t))) = 0$, one has

$$\Omega_\infty(S(\infty)) = \bigcup_{k\in W} \{\mathbf{w_k}(\infty)\}. \tag{52}$$

Obviously, if "impulses" exist, there exists some $W \neq \{1, \cdots, N\}$, such that $\mathcal{P}(E'_{W,2}) \neq 0$ in (47). Consequently, we cannot conclude $\mathcal{P}(E_{\{1,\cdots,N\}}) = 1$.

Theorem 1 guarantees that two existing instances, namely, the FSCL1 [10] and the DSCL [14], do not have neuron underutilization problem. However, our theorem is not applicable to two other instances: the FSCL2 and the DECL. In the FSCL2, $M_i(S)$ is time-dependent. In the DECL, we have $F_i(\mathbf{x}, S) = (b_i/\sum_j b_j)\, d(\mathbf{x}, \mathbf{w_i})$ such that when $b_i < \infty$ and $\sum_j b_j = \infty$ in the limit when $t \to \infty$, we have $F_i(\mathbf{x}, S) = 0$, which violates the requirement on $F_i(\mathbf{x}, S)$ [Condition 4b) in Theorem 1].

Hence, interesting future works along this direction of research will be in relaxing the requirements on the bias-updating function and the smoothness requirement on $p(\mathbf{x})$.

IV. Conclusions

In this paper, we have analyzed a class of CL models called the MBCL model, which can be efficiently implemented in sequential environment. We have proved that under suitable conditions that the MBCL can avoid neuron underutilization problem with probability one, according to Theorem 1. Indeed, as we have summarized and analyzed in Section II, neuron underutilization implies suboptimal performance (e.g., in classification, vector quantizer design, and probability density function estimation) and sensitivity to initial conditions, and hence must be avoided in unsupervised learning. Furthermore, by proving that an MBCL can avoid neuron underutilization with probability one, the convergence property of the MBCL can be arrived at by using the convergence theorem from Kosmatopoulos and Christodoulou [20].

Conventionally, almost all theoretical studies on existing CL models have been concentrating on finding the energy function that a particular CL model is minimizing. Although it is possible to demonstrate the global optimality of these approaches, we have no idea on whether neuron underutilization problem will occur. For example, although in case of the SCL, it has a well-defined energy function (e.g., in [1]) indicating that it is possible to converge to a globally optimal solution, the effect of poor initialization cannot be deduced. On the other hand, the analysis of "neural-gas" by Martinetz *et al.* [12] indicates that the distribution of weight vectors approaches the codevectors distribution of a globally optimal vector quantizer under the limiting conditions of large N and sufficiently smooth $p(\mathbf{x})$. This indeed could imply that no neuron is underused, although this is true only in the limit of large N and smooth $p(\mathbf{x})$. However, in this paper, we have presented a qualitative analysis on the ability of the MBCL to avoid neuron underutilization problem. To the best of our knowledge, we are the first to consider this problem in a mathematically rigorous way.

By proposing the theorem, we define sufficient conditions for an MBCL to avoid neuron underutilization. We hope that new instances of the MBCL can be suggested based on this framework, such that some optimality criteria can be achieved specific to the targeted applications, and yet neuron underutilization problem can be avoided. Indeed, if neuron underutilization problem will lead to suboptimal performances in the targeted applications, we will expect this new MBCL to have good performance. Note that the major difference between existing MBCL's to which our theorem is applicable differs only in the way bias factors are increased, i.e., $F_k(\mathbf{x}, S)$ in (20b). We expect that by modifying this function, specific optimality criteria can be achieved. For example, we have recently proposed an instance of the MBCL, the DSCL [14], which can satisfy the equidistortion principle in vector quantizer design [30] and outperforms a number of existing CL models in vector quantizer design. We expect that this is partially due to its capability in avoiding neuron underutilization problem.

APPENDIX A
PROOF OF LEMMA 1:

With reference to (22), $\Omega_i(S)$ at any state S is constructed by the intersection of $N-1$ sets Γ_{ij}. Note that for any two sets A_1 and A_2, we have

$$|D_{A_1}| < \infty \wedge |D_{A_2}| < \infty \Rightarrow |D_{A_1 \cap A_2}| < \infty. \tag{53}$$

Hence, if $|D_{\Gamma_{ij}}| < \infty \, \forall \, i, \, j$, we can deduce that $|D_{\Omega_i(S)}| < \infty$.

Let us define $\Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$, where

$$\Gamma(\mathbf{c_1}, \mathbf{c_2}, q) = \{\mathbf{x} \in \Re^K : d(\mathbf{x}, \mathbf{c_1}) < q \, d(\mathbf{x}, \mathbf{c_2})\} \quad \text{where } 0 \le q \le 1. \tag{54}$$

Furthermore, let us define $q_{ij} = M_i(S)/M_j(S)$. Then, for a particular set Γ_{ij} defined in (22), if $0 \le q_{ji} \le 1$, we have $\Gamma_{ij} = \Gamma(\mathbf{w_i}, \mathbf{w_j}, q_{ji})$. Otherwise, we can express Γ_{ij} as $\Re^K \backslash \Gamma(\mathbf{w_j}, \mathbf{w_i}, q_{ij})$. In the former case, if $\Gamma(\mathbf{w_i}, \mathbf{w_j}, q_{ji})$ is connected, it must be true that Γ_{ij} is connected as well. In the latter case, since $\Re^K$ is obviously connected, the connectedness of $\Gamma(\mathbf{w_j}, \mathbf{w_i}, q_{ij})$ implies that $|D_{\Gamma_{ij}}| < \infty$. In either case, we have to prove the connectedness of $\Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$.

In order to prove the connectedness of $\Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$, we consider a point $\mathbf{x} \in \Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$, and define the line segment $L(\mathbf{x})$: $\mathbf{z} = \beta(\mathbf{x} - \mathbf{c_1}) + \mathbf{c_1} \; \forall \beta \in [0, 1]$. If any point $\mathbf{z}$ on this line segment $L(\mathbf{x})$ is also within the set $\Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$, we know that $\mathbf{x}$ is connected with $\mathbf{c_1}$. Consequently, for any two points $\mathbf{x_1}, \mathbf{x_2} \in \Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$, they are connected through the point $\mathbf{c_1}$, which means that $\Gamma(\mathbf{c_1}, \mathbf{c_2}, q)$ is connected. Hence, we have to prove

$$\forall q \in [0, 1], \; \forall \mathbf{x} \in \Re^K, \; d(\mathbf{x}, \mathbf{c_1}) \le q \, d(\mathbf{x}, \mathbf{c_2}) \Rightarrow \forall \mathbf{z} \in L(\mathbf{x}), \; d(\mathbf{z}, \mathbf{c_1}) \le q \, d(\mathbf{z}, \mathbf{c_2}). \tag{55}$$

Let $q \in (0, 1]$, then

$$\begin{aligned} &\forall \mathbf{x} \in \Re^K, \; d(\mathbf{x}, \mathbf{c_1}) \le q \, d(\mathbf{x}, \mathbf{c_2}) \\ &\quad \Leftrightarrow G(\mathbf{x} - \mathbf{c_1}) \le G(f(q)(\mathbf{x} - \mathbf{c_2})) \\ &\qquad (\text{we define } g(f(q)) = q \text{ where } g(\cdot) \text{ is defined in } (19)) \\ &\quad \Leftrightarrow ||\mathbf{x} - \mathbf{c_1}|| \le f(q)||\mathbf{x} - \mathbf{c_2}|| \\ &\qquad (\text{according to } (19) \text{ and } ||a \cdot \mathbf{u}|| = a||\mathbf{u}|| \\ &\qquad \text{for any } a > 0 \text{ since } ||\cdot|| \text{ is a norm}). \end{aligned} \tag{56}$$

Note that since the function $g(\cdot)$ in (19) is strictly increasing, its inverse $f(\cdot)$ must exist, and that $0 \le f(q) \le 1 \; \forall q \in [0, 1]$. Now, let us evaluate the following difference:

$$\begin{aligned} &||\mathbf{z} - \mathbf{c_1}|| - f(q)||\mathbf{z} - \mathbf{c_2}|| \\ &\quad = \beta||\mathbf{x} - \mathbf{c_1}|| - f(q)||\mathbf{z} - \mathbf{c_2}|| \; (\text{since } ||\cdot|| \text{ is a norm}) \\ &\quad \le \beta||\mathbf{x} - \mathbf{c_1}|| + f(q)||\mathbf{x} - \mathbf{z}|| - f(q)||\mathbf{x} - \mathbf{c_2}|| \\ &\qquad (\text{since } ||\mathbf{x} - \mathbf{c_2}|| \le ||\mathbf{x} - \mathbf{z}|| + ||\mathbf{z} - \mathbf{c_2}||) \\ &\quad \le \beta||\mathbf{x} - \mathbf{c_1}|| + f(q)(1 - \beta)||\mathbf{x} - \mathbf{c_1}|| - ||\mathbf{x} - \mathbf{c_1}|| \\ &\qquad (\text{since } \mathbf{x} \in \Gamma \Rightarrow ||\mathbf{x} - \mathbf{c_1}|| \le f(q)||\mathbf{x} - \mathbf{c_2}|| \\ &\qquad \text{according to } (56)) \\ &\quad = (1 - \beta)(f(q) - 1)||\mathbf{x} - \mathbf{c_1}|| \le 0. \end{aligned} \tag{57}$$

Hence

$$\begin{aligned} &||\mathbf{x} - \mathbf{c_1}|| \le f(q)||\mathbf{x} - \mathbf{c_2}|| \\ &\quad \Rightarrow ||\mathbf{z} - \mathbf{c_1}|| \le f(q)||\mathbf{z} - \mathbf{c_2}|| \\ &\quad \Leftrightarrow G(\mathbf{z} - \mathbf{c_1}) \le qG(\mathbf{z} - \mathbf{c_2}) \\ &\quad \Leftrightarrow d(\mathbf{z}, \mathbf{c_1}) \le q \, d(\mathbf{z}, \mathbf{c_2}). \end{aligned} \tag{58}$$

This proves (55), which completes our proof. □

APPENDIX B
BOUNDEDNESS OF NORMALIZED BIAS FACTORS

In the following, we will show that the normalized bias factors defined in (48) are bounded $\forall t > 0$.

Proof: Let us consider the updating equation of the normalized bias factor c_i in (51). Recalled that $F_i(\mathbf{x}, S)$ is nonnegative and bounded [Condition 4a) of Theorem 1], and that $\gamma_i(\mathbf{x}, S)$ is either zero or one.

In order to prove the boundedness of normalized bias factors, we assume that for the ith neuron, it is true that $c_i \to \infty$ as $t \to \infty$. Then

$$\exists t' \text{ s.t. } \forall t > t', \; \mathcal{P}(c_i(t) > C_F) = 1 \tag{59}$$

where C_F is the upper bound of $F_i(\mathbf{x}, S)$ as defined in Condition 4a (Theorem 1). Note that the following is true:

$$c_i - \gamma_i(\mathbf{x}, S)F_i(\mathbf{x}, S) \ge c_i - F_i(\mathbf{x}, S) \ge c_i - C_F. \tag{60}$$

By combining (59) and (60) with (51), we have

$$\forall t > t', \; \mathcal{P}(\triangle c_i(t) < 0) = 1. \tag{61}$$

This implies that $\forall t > t', \; \mathcal{P}(c_i(t) \le c_i(t')) = 1$, which contradicts our original assumption that $c_i \to \infty$ as $t \to \infty$.

This completes our proof. □

REFERENCES

[1] T. Kohonen, *Self-Organizing Maps.* New York: Springer-Verlag, 1995.

[2] ———, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, pp. 59–69, 1982.

[3] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[4] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[5] S. Grossberg, "Adaptive pattern classification and universal recoding; II. Feedback, expectation, olfaction, illusions," *Biol. Cybern.*, vol. 23, pp. 187–202, 1976.

[6] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognitive Sci.*, vol. 9, pp. 75–112, 1985.

[7] G. A. Carpenter and S. Grossberg, "ART2: Self-organization of stable category recognition codes for analog input patterns," *Appl. Opt. 1*, vol. 26, no. 23, pp. 4919–4930, Dec. 1987.

[8] N. R. Pal, J. C. Bezdek, and E. C. K. Tsao, "Generalized clustering networks and Kohonen's self-organizing scheme," *IEEE Trans. Neural Networks*, vol. 4, pp. 549–557, July 1993.

[9] N. Karayiannis, J. C. Bezdek, N. R. Pal, R. J. Hathaway, and P.-I. Pai, "Repairs to GLVQ: A new family of competitive learning schemes," *IEEE Trans. Neural Networks*, vol. 7, pp. 1062–1071, Sept. 1996.

[10] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton, "Competitive learning algorithms for vector quantization," *Neural Networks*, vol. 3, pp. 277–290, 1990.

[11] A. K. Krishnamurthy, S. C. Ahalt, D. E. Melton, and P. Chen, "Neural networks for vector quantization of speech and images," *IEEE J. Select. Areas Commun.*, vol. 8, pp. 1449–1457, Oct. 1990.
[12] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten, "Neural-gas network for vector quantization and its application to time-series prediction," *IEEE Trans. Neural Networks*, vol. 4, pp. 558–569, July 1993.
[13] D. Butler and J. Jiang, "Distortion equalized fuzzy competitive learning for image data vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'96)*, London, U.K., May 1996, pp. 3390–3393.
[14] C. S.-T. Choy and W.-C. Siu, "Distortion sensitive competitive learning for vector quantizer design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'97)*, Munich, Germany, Apr. 1997, pp. 3405–3408.
[15] D. DeSieno, "Adding a conscience to competitive learning," in *Proc. IEEE Int. Conf. Neural Networks*, 1988, vol. 1, pp. 117–124.
[16] L.-H. Chen and S. Chang, "An adaptive conscientious competitive learning algorithm and its applications," *Pattern Recognitions*, vol. 27, no. 12, pp. 1787–1813, 1994.
[17] S. Grossberg, "On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks," *J. Statist. Phys.*, vol. 1, no. 2, pp. 319–350, 1969.
[18] ———, "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors," *Biol. Cybern.*, vol. 23, pp. 121–134, 1976.
[19] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[20] E. B. Kosmatopoulos and M. A. Christodoulou, "Convergence properties of a class of learning vector quantization algorithms," *IEEE Trans. Image Processing*, vol. 5, pp. 361–368, Feb. 1996.
[21] A. I. Gonzalez, M. Grana, and A. Dànjou, "An analysis of the GLVQ algorithm," *IEEE Trans. Neural Networks*, vol. 6, pp. 1012–1016, July 1995.
[22] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, pp. 4–29, Apr. 1984.
[23] D. M. Clark and K. Ravishankar, "A convergence theorem for Grossberg learning," *Neural Networks*, vol. 3, pp. 87–92, 1990.
[24] J. A. Kangas, T. K. Kohonen, and J. T. Laaksonen, "Variants of self-organizing maps," *IEEE Trans. Neural Networks*, vol. 1, pp. 93–99, Mar. 1990.
[25] L. A. Zadeh, "Fuzzy sets," *Inform. Contr.*, vol. 8, pp. 338–353, 1965.
[26] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.
[27] F. L. Chung and T. Lee, "Fuzzy competitive learning," *Neural Networks*, vol. 7, no. 3, pp. 539–551, 1994.
[28] C. Zhu, L. H. Li, T. J. Wang, and Z. Y. He, "Partial-distortion-weighted fuzzy competitive learning algorithm for vector quantization," *Electron. Lett.*, vol. 30, no. 6, pp. 505–506, Mar. 1994.
[29] N. B. Karayiannis and P.-I. Pai, "Fuzzy algorithms for learning vector quantization," *IEEE Trans. Neural Networks*, vol. 7, pp. 1196–1211, Sept. 1996.
[30] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. 25, pp. 373–380, July 1979.

Clifford Sze-Tsan Choy received the B.Eng. degree with first honors in electronic engineering in 1992 and the Ph.D. degree in electronic engineering in 1998, both from the Hong Kong Polytechnic University, Hong Kong.

He is now working as a Senior Systems Engineer at the UNI-Tech Technologies Limited. His research interests include neural networks, optimization, and pattern recognition.

Wan-Chi Siu (S'77–M'77–SM'90) received the Associateship degree from the Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic), the M.Phil. degree from the Chinese University of Hong Kong, and the Ph.D. degree from the Imperial College of Science, Technology and Medicine, London, in 1975, 1977, and 1984, respectively.

He was with the Chinese University of Hong Kong between 1975 and 1980. He then joined the Hong Kong Polytechnic University in 1980 and became Chair Professor and Associate Dean of Engineering Faculty in 1992. He has been Chair Professor and Head of Department of Electronic and Information Engineering of the same university since 1994. He has published more than 160 research papers. His research interests include digital signal processing, fast computational algorithms, transforms, video coding, computational aspects of image processing and pattern recognition, and neural networks.

Dr. Siu is a Member of the Editorial Board of the *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, an honorary advisor of the Editorial Board of the *Journal of Data Acquisition and Processing in China*, and an overseas member of the Editorial Board of the *IEE Review*. He is also a Guest Editor of the Special Issue on ISCAS'97 of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II, published in May 1998. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1995 to 1997. He was the General Chairman of the International Symposium on Neural Networks, Image, and Speech Processing (ISSIPNN'94), and a Cochair of the Technical Program Committee of the IEEE International Symposium on Circuits and Systems (ISCAS'97) which were held in Hong Kong in April 1994 and June 1997, respectively. Between 1991 and 1995, he was a member of the Physical Sciences and Engineering Panel of the Research Grants Council (RGC), Hong Kong Government, and in 1994 he chaired the first Engineering and Information Technology Panel to assess the research quality of 19 Cost Centers (departments) from all universities in Hong Kong. He is a Chartered Engineer, a Fellow of the Institute of Electrical Engineers and the HKIE, and has also been listed in Marquis *Who's Who in the World*, Marquis *Who's Who in Science, and Engineering*, and other citation biographies.