Learning of Associative Memory Networks by Penalty Methods

Niijima, Koichi
Department of Control Engineering and Science Kyushu Institute of Technology

出版情報：RIFIS Technical Report. 113, 1995-04-26. Research Institute of Fundamental Information Science, Kyushu University
バージョン：
権利関係：

RIFIS-TR-CS-113

RIFIS Technical Report

Learning of Associative Memory Networks by Penalty Methods

Koichi Niijima

April 26, 1995

Research Institute of Fundamental Information Science
Kyushu University 33
Fukuoka 812, Japan
E-mail: niijima@ces.kyutech.ac.jp

Learning of Associative Memory Networks by Penalty Methods

Koichi Niijima

Abstract– **This paper concerns the learning of associative memory networks. We derive inequality associative conditions for stored patterns in the network. Under these associative conditions, we find regions each of which is mapped by the network function into a neighbor of an associative pattern. To make large the regions, a functional is derived using their shape. The functional is minimized under the inequality associative conditions. We show that this minimization problem has a unique solution, and solve the problem by combining the penalty methods with the gradient methods. This solving process gives a learning algorithm for associative networks. Our theory is first used to analyze two-layer autoassociative networks. It is shown that the network function becomes a contraction mapping in each of the regions derived under inequality autoassociative conditions. We also show that the function has a fixed point extremely near a stored pattern. This implies that the region obtained is a domain of attraction and that the fixed point is its attractor. Next, our learning algorithm is applied to make a heteroassociative network which is useful for solving classification problems. By adding one more layer to the network, we construct a three-layer autoassociative network whose input-output function is shown to be a contraction mapping in some domains. In simulations, efficiency of our two autoassociative networks is verified in character recognition.**

I. Introduction

Associative memory networks have been studied from mainly two aspects. One of them is research on learning algorithms for these networks and the other concerns capability of such networks. There are several learning algorithms of associative memories such as the correlation recording, the generalized-inverse recording [6], and the Ho-Kashyap algorithm [5]. In [4], these learning techniques have been surveyed together with the capacity and performance of associative memories. Recently, Caianiello and Benedictis [2] proposed a memorization rule for associative memories with minimum connectivity. Storage capacity of associative memory networks has been investigated by Amari [1], Cottrell [3], McEliece, Posner, Rodemich, and Venkatesh [7], from the viewpoint of domains of attraction. These papers on storage capacity, however, treat only binary-valued vectors.

In our recent works [8] and [9], we proposed a learning method for associative memory networks. The method can deal with analog patterns and is based on the domains of attraction in the network. These domains were derived under equality associative conditions. Associative conditions are not necessarily of equality type. Equality associative conditions restrict the number of stored patterns and the size of domains of attraction.

In this paper, we consider inequality associative conditions in place of equality ones for stored patterns. Under these relaxed conditions, we derive regions including stored patterns, each of which is mapped by the network function into a neighbor of an associative pattern. It is shown that the region is larger than that obtained under equality associative conditions. We also show that these regions are mutually disjoint. These results are given in Section II.

We wish to determine connection weights in the network so that the region obtained becomes as large as possible. For the purpose, a functional to be minimized is derived using the shape of the region. This minimization problem, however, does not have a unique threshold parameter which is a part of weights in the network. So we impose some conditions on the threshold parameters. Thus a modified functional is derived and it is minimized subject to the inequality associative constraints. This minimization problem has a unique solution, and can be solved by combining the penalty methods with the gradient methods. This solving process gives a learning algorithm for determining connection weights of associative networks. These are described in Section III.

Our theory is first used to analyze two-layer autoassociative networks. We put on the network the condition that when a binary valued stored pattern is input in the network, almost the same pattern is output. This condition is expressed in an inequality form and the above theory is applied to obtain regions each of which is mapped by the network function to almost a stored pattern. We show by the contraction mapping theorem that the function has a unique fixed point in the region, which is extremely near a stored pattern. This implies that the region is a domain of attraction and that the fixed point is its attractor. A minimization problem to determine connection weights of the network is easily derived using the method described above.

K. Niijima is with the Department of Control Engineering and Science, Kyushu Institute of Technology, Iizuka 820, JAPAN.

These results are described in Section IV.

Our learning algorithm is also applied to make a specified heteroassociative network in which only one output neuron fires for one stored pattern. Neural networks of this type are useful for solving classification problems. By adding one more layer to the network, we construct a three-layer autoassociative network. It is shown that the network function becomes a contraction mapping in some domains. These are discussed in Section V.

In simulations which will be given in Section VI, character recognition ability of our two autoassociative networks is compared with that of the networks obtained under equality associative conditions in [8] and [9].

Concluding remarks are described in Section VII.

II. Inequality Associative Conditions and Neighbors of Stored Patterns

Let n be the number of input nodes and ℓ the number of output neuron units. We consider the following neural network:

$$y_i = f(\sum_{j=1}^{n} w_{ij}x_j - \theta_i), \qquad i = 1, 2, ..., \ell, \tag{1}$$

where w_{ij} and θ_i denote weights between the input and output layers. The function $f(t)$ indicates the sigmoid function

$$f(t) = \frac{1}{1 + \exp(-t)}.$$

We put $W_i = {}^t(w_{i1}, w_{i2}, ..., w_{in})$, $x = {}^t(x_1, x_2, ..., x_n)$ with the transpose symbol ^{t} and rewrite (1) as

$$\begin{aligned} y_i &= f(W_i \cdot x - \theta_i) \\ &\equiv \varphi_i(x), \qquad i = 1, 2, ..., \ell, \end{aligned} \tag{2}$$

where $\cdot$ denotes the inner product symbol. We also put $\varphi(x) = {}^t(\varphi_1(x), \varphi_2(x), ..., \varphi_\ell(x))$ and $y = {}^t(y_1, y_2, ..., y_\ell)$. Then (2) may be written as

$$y = \varphi(x). \tag{3}$$

Let x^ν,$\nu = 1, 2, ..., m$, denote patterns to be stored in the network. We assume that when x^ν is input into (2), a number larger than $1-\varepsilon$ or smaller than ε is output, where ε is a sufficiently small positive parameter. We define two sets of indexes:

$$I_{i,+} = \{\nu \mid f(W_i \cdot x^\nu - \theta_i) \geq 1 - \varepsilon\}, \tag{4}$$

$$I_{i,-} = \{\nu \mid f(W_i \cdot x^\nu - \theta_i) \leq \varepsilon\}. \tag{5}$$

Using the monotonicity of $f(t)$ and the inverse function $t = \ln(s/(1-s))$ of $s = f(t)$, we have

$$V_i \cdot x^\nu - \eta_i \geq 1 \qquad \text{for} \quad \nu \in I_{i,+}, \tag{6}$$

where $W_i = V_i \ln((1-\varepsilon)/\varepsilon)$ and $\theta_i = \eta_i \ln((1-\varepsilon)/\varepsilon)$. On the other hand, we have

$$V_i \cdot x^\nu - \eta_i \leq -1 \qquad \text{for} \quad \nu \in I_{i,-}. \tag{7}$$

We call (6) and (7) inequality associative conditions for the network (2). Under the conditions (6) and (7), the following theorem holds.

Theorem 1: Suppose that V_i and η_i satisfy (6) and (7). For $0 < \rho < 1$ and each stored pattern x^ν, we define the region $D_\rho(x^\nu)$ in R^n by

$$D_\rho(x^\nu) = \{x \mid |V_i \cdot (x - x^\nu)| \leq \rho |V_i \cdot x^\nu - \eta_i|, \ i = 1, 2, ..., \ell\}.$$

Then we have for any x, $\widetilde{x} \in D_\rho(x^\nu)$,

$$|\varphi_i(x) - \varphi_i(\widetilde{x})| \leq \varepsilon^{1-\rho} \ln\frac{1}{\varepsilon} |V_i \cdot (x - \widetilde{x})|, \quad i = 1, 2, ..., \ell. \tag{8}$$

Especially, when $\widetilde{x} = x^\nu$, we have

$$\begin{aligned} |\varphi_i(x) - \varphi_i(x^\nu)| &\leq \rho\, \varepsilon^{1-\rho} \ln\frac{1}{\varepsilon} |V_i \cdot x^\nu - \eta_i| \\ &\equiv \varepsilon_i^\nu, \qquad i = 1, 2, ..., \ell. \end{aligned} \tag{9}$$

Proof: See Appendix. □

In [8] and [9], we put the equality output condition

$$f(W_i \cdot x^\nu - \theta_i) = 1 - \varepsilon \quad \text{or} \quad \varepsilon$$

which is stronger than the condition in (4) and (5). Then the equality holds in (6) and (7) and we have $|V_i \cdot x^\nu - \eta_i| = 1$. Under (6) and (7), we have $|V_i \cdot x^\nu - \eta_i| \geq 1$. Therefore, the region $D_\rho(x^\nu)$ is larger than that obtained under equality associative conditions.

The right hand side ε_i^ν of (9) is small enough for a sufficiently small ε. This implies that any pattern in $D_\rho(x^\nu)$ is recognized as the pattern x^ν.

Assume that for any pair (x^ν, x^μ), $\nu \neq \mu$, there exists at least one index i_0 such that if $\nu \in I_{i_0,+}$, then $\mu \in I_{i_0,-}$ holds. That is,

$$\varphi_{i_0}(x^\nu) = f(W_{i_0} \cdot x^\nu - \eta_{i_0}) \geq 1 - \varepsilon, \tag{10}$$

$$\varphi_{i_0}(x^\mu) = f(W_{i_0} \cdot x^\mu - \eta_{i_0}) \leq \varepsilon. \tag{11}$$

Then it can be shown that $D_\rho(x^\nu), \nu = 1, 2, ..., m$ are mutually disjoint. The proof is as follows: Assume $D_\rho(x^\nu) \cap D_\rho(x^\mu) \neq \phi$, where ϕ indicates the empty set. Then there exists $x^* \in D_\rho(x^\nu) \cap D_\rho(x^\mu)$. By (9) of Theorem 1, we have

$$|\varphi_{i_0}(x^*) - \varphi_{i_0}(x^\nu)| \leq \varepsilon_{i_0}^\nu, \tag{12}$$

$$|\varphi_{i_0}(x^*) - \varphi_{i_0}(x^\mu)| \leq \varepsilon_{i_0}^\mu. \tag{13}$$

Combining (10) with (12), we get

$$\varphi_{i_0}(x^*) \geq 1 - \varepsilon - \varepsilon_{i_0}^\nu. \tag{14}$$

On the other hand, we have by (11) and (13),

$$\varphi_{i_0}(x^*) \leq \varepsilon + \varepsilon_{i_0}^\mu.$$

This contradicts (14).

III. Learning Method

From the viewpoint of recognition ability, it is desirable for the region $D_\rho(x^\nu)$ in Theorem 1 to be as large as possible. One way to make large $D_\rho(x^\nu)$ is to maximize the distances from x^ν to the hyperplanes $H_i : V_i \cdot (x - x^\nu) = \pm\rho$ $|V_i \cdot x^\nu - \eta_i|$, $i = 1, 2, ..., n$. The distance is sought as follows: Since a normal vector for H_i is V_i, a vector x from x^ν with the direction V_i is represented as $x = x^\nu + \alpha V_i$. By the condition that this x is on H_i, we obtain $\alpha = \pm\rho|V_i \cdot x^\nu - \eta_i|/\|V_i\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. Therefore, the distance is given by

$$\|x - x^\nu\| = \frac{\rho|V_i \cdot x^\nu - \eta_i|}{\|V_i\|}.$$

We wish to determine the weights V_i and η_i so that the distance is maximized under the conditions (6) and (7). This leads to the following minimization problem:

$$\frac{\|V_i\|^2}{(V_i \cdot x^\nu - \eta_i)^2} \longrightarrow \min$$

subject to (6) and (7). However, the solutions V_i and η_i depend on the index ν of the stored pattern x^ν. So, using the inequality

$$\frac{\|V_i\|}{|V_i \cdot x^\nu - \eta_i|} \leq \|V_i\| \tag{15}$$

which follows from (6) and (7), we minimize

$$\|V_i\|^2$$

subject to (6) and (7). This minimization problem can be solved by the penalty methods. Let C be a sufficiently large penalty constant and introduce the functional

$$J_i^0 = \|V_i\|^2 + C \Big(\sum_{\nu \in I_{i,+}} (1 - V_i \cdot x^\nu + \eta_i)_+^2 + \sum_{\mu \in I_{i,-}} (1 + V_i \cdot x^\mu - \eta_i)_+^2 \Big).$$

The functional J_i^0 is strictly convex with respect to V_i. On the variable η_i, the functional J_i^0 is convex, but not strictly convex. Therefore, the solutions of this problem are not unique. To obtain a functional yielding a unique minimum, we return to to the functional (15).

For the index i such that $I_{i,+} \neq \phi$ and $I_{i,-} \neq \phi$, we choose $\nu \in I_{i,+}$ and $\mu \in I_{i,-}$. Then we have from (6) and (7),

$$V_i \cdot x^\mu + 1 \leq \eta_i \leq V_i \cdot x^\nu - 1.$$

Put $\eta_i^{\nu,\mu} = V_i \cdot (x^\nu + x^\mu)/2$. Then $|V_i \cdot x^\nu - \eta_i^{\nu,\mu}| = |V_i \cdot x^\mu - \eta_i^{\nu,\mu}|$ holds. Using these $\eta_i^{\nu,\mu}$, we determine η_i so as to minimize the functional

$$\sum_{\nu \in I_{i,+}} \sum_{\mu \in I_{i,-}} (\eta_i^{\nu,\mu} - \eta_i)^2.$$

This makes $|V_i \cdot x^\nu - \eta_i|$ for all ν keep as equally as possible, and justifies our idea that minimizes $\|V_i\|^2$ instead of $\|V_i\|^2/(V_i \cdot x^\nu - \eta_i)^2$. We thus arrive at a minimization problem:

$$\|V_i\|^2 + \sum_{\nu \in I_{i,+}} \sum_{\nu \in I_{i,-}} (\eta_i^{\nu,\mu} - \eta_i)^2 \longrightarrow \min \tag{16}$$

subject to (6) and (7). We solve the problem (16) by a penalty method:

$$J_i = \|V_i\|^2 + \sum_{\nu \in I_{i,+}} \sum_{\mu \in I_{i,-}} (\eta_i^{\nu,\mu} - \eta_i)^2 + C\Big(\sum_{\nu \in I_{i,+}} (1 - V_i \cdot x^\nu + \eta_i)_+^2 + \sum_{\mu \in I_{i,-}} (1 + V_i \cdot x^\mu - \eta_i)_+^2 \Big) \longrightarrow \min. \tag{17}$$

Since the functional J_i is strictly convex with respect to V_i and η_i, it possesses a unique minimum. This minimization problem can be solved by various gradient methods.

For the index i such that $I_{i,+} = \phi$, we choose $\eta_i = 1$. Then J_i^0 has a unique minimum $V_i = 0$.

For the index i such that $I_{i,-} = \phi$, we choose $\eta_i = -1$. Then J_i^0 also has a unique minimum $V_i = 0$.

Therefore, it suffices to solve the minimization problem (17) only for i such that $I_{i,+} \neq \phi$ and $I_{i,-} \neq \phi$. Thus we can compute the weights of the network. The computing process gives our learning algorithm for associative memory networks.

IV. Two-layer Autoassociative Memory Network

We apply our theory to analyze two-layer autoassociative memory networks. The stored patterns x^ν,$\nu = 1, 2, ..., m$, are assumed to be $(0,1)-$valued vectors. We choose $\ell = n$ in the network (2). As an output condition for the stored pattern $x^\nu = {}^t(x_1^\nu, x_2^\nu, ..., x_n^\nu)$, we use

$$f(W_i \cdot x^\nu - \theta_i) \begin{cases} \geq 1 - \varepsilon & \text{for } x_i^\nu = 1, \\ \leq \varepsilon & \text{for } x_i^\nu = 0. \end{cases} \tag{18}$$

The condition (18) means that when x^ν is input, a pattern extremely near x^ν is output, because of a sufficiently small parameter ε. Namely, the network satisfying (18) is almost of autoassociative type. The condition (18) may be rewritten as

$$V_i \cdot x^\nu - \eta_i \geq 1 \quad \text{for } x_i^\nu = 1, \tag{19}$$

$$V_i \cdot x^\nu - \eta_i \leq -1 \quad \text{for } x_i^\nu = 0, \tag{20}$$

which are called inequality autoassociative conditions. Theorem 1 holds for such a network and the regions $D_\rho(x^\nu)$,

$\nu = 1, 2, ..., m$, are mutually disjoint. The region $D_\rho(x^\nu)$ is larger than that obtained under equality autoassociative conditions in [8].

From Theorem 1, we obtain the following theorem.

Theorem 2: Suppose that V_i and η_i satisfy (19) and (20). Then we have for any x, $\widetilde{x} \in D_\rho(x^\nu)$,

$$\|\varphi(x) - \varphi(\widetilde{x})\| \leq \kappa \|x - \widetilde{x}\|, \tag{21}$$

where

$$\kappa = \sqrt{\sum_{i=1}^{n} \|V_i\|^2}\ \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}.$$

If ε is sufficiently small so as to satisfy $\kappa < 1$, the function φ becomes a contraction mapping in $D_\rho(x^\nu)$.

Moreover, the equation $x = \varphi(x)$ has a unique solution $x^{\nu,*}$ in $D_\rho(x^\nu)$ satisfying $\|x^{\nu,*} - x^\nu\| \leq K\varepsilon$ with a positive constant K.

Proof: The inequality (21) is proved by estimating (8) in Theorem 1 from above. Indeed,

$$|\varphi_i(x) - \varphi_i(\widetilde{x})| \leq \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}\ \|V_i\|\ \|x - \widetilde{x}\|.$$

Summing up both sides after squared, we obtain (21).

The latter can be proved by the contraction mapping theorem (See p.65 of [10]). To apply this theorem, let us define a ball $B_\delta(x^\nu)$ in $D_\rho(x^\nu)$ by

$$B_\delta(x^\nu) = \{x \mid \|x - x^\nu\| \leq \delta,\ \ \delta = \min_{i=1,2,...,n} \frac{\rho}{\|V_i\|}\ \}$$

and put

$$r = \frac{\|\varphi(x^\nu) - x^\nu\|}{1 - \kappa}.$$

By the contraction mapping theorem, φ has a fixed point $x^{\nu,*}$ in $B_\delta(x^\nu)$, that is, there exists a solution $x^{\nu,*} \in B_\delta(x^\nu)$ such that $x^{\nu,*} = \varphi(x^{\nu,*})$. Furthermore, we have

$$\|x^{\nu,*} - x^\nu\| \leq r.$$

Since $r \leq K\varepsilon$ holds for a positive constant K, we obtain the second result of Theorem 2. □

Let $\{x^{(k)}\}_{k=0,1,...}$ be a sequence generated by the iteration

$$x^{(k+1)} = \varphi(x^{(k)}), \quad k = 0, 1, ...,$$

where $x^{(0)}$ is in $D_\rho(x^\nu)$. This sequence converges to the fixed point $x^{\nu,*}$. Therefore, we may say $D_\rho(x^\nu)$ a domain of attraction.

Let φ^0 be the identity mapping in R^n and define φ^q, $q \geq 1$, by $\varphi^q(x) = \varphi(\varphi^{q-1}(x))$ recursively. We consider the domains

$$D_\rho^q(x^{\nu,*}) = \{x \mid |V_i \cdot (\varphi^q(x) - x^{\nu,*})| \leq \rho |V_i \cdot x^{\nu,*} - \eta_i|,$$
$$i = 1, 2, ..., n\ \}. \tag{22}$$

Since $x^{\nu,*}$ is extremely near x^ν, the domain $D_\rho^0(x^{\nu,*})$ is almost equal to the domain $D_\rho(x^\nu)$. We also have the following theorem.

Theorem 3: Suppose that ε satisfies the assumption of Theorem 2. Then the function φ becomes a contraction mapping in $D_\rho^0(x^{\nu,*})$.

Put

$$L_\nu = \sqrt{\sum_{i=1}^{n} (V_i \cdot x^{\nu,*} - \eta_i)^2}.$$

Then we have for any $x \in D_\rho^q(x^{\nu,*})$,

$$\|\varphi^{q+1}(x) - x^{\nu,*}\| \leq L_\nu \rho\ \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}. \tag{23}$$

Moreover, if ε is small enough so as to satisfy

$$L_\nu \max_{i=1,2,...,n} \|V_i\|\ \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon} \leq 1,$$

we have the inclusions

$$D_\rho(x^\nu) \cong D_\rho^0(x^{\nu,*}) \subset D_\rho^1(x^{\nu,*}) \subset D_\rho^2(x^{\nu,*}) \subset ... \tag{24}$$

Proof: In the same way as in Theorem 1, we can show that for any x, $\widetilde{x} \in D_\rho^0(x^{\nu,*})$,

$$|\varphi_i(x) - \varphi_i(\widetilde{x})| \leq \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}\ |V_i \cdot (x - \widetilde{x})|, \quad i = 1, 2, ..., n. \tag{25}$$

By estimating this inequality, we obtain the same inequality as (21) which implies that φ is a contraction mapping also in $D_\rho^0(x^{\nu,*})$.

The proof of (23) is as follows: For any $x \in D_\rho^q(x^{\nu,*})$, we have $u \equiv \varphi^q(x) \in D_\rho^0(x^{\nu,*})$. Hence, from (25) and (22),

$$\begin{aligned} |\varphi_i(u) - \varphi_i(x^{\nu,*})| &\leq \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}\ |V_i \cdot (u - x^{\nu,*})| \\ &\leq \rho\ \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}\ |V_i \cdot x^{\nu,*} - \eta_i|. \end{aligned}$$

Summing up both sides after squared, we obtain (23).

The inclusions (24) are proved as follows: For any $x \in D_\rho^q(x^{\nu,*})$, we have from (23),

$$\begin{aligned} |V_i \cdot (\varphi^{q+1}(x) - x^{\nu,*})| &\leq \|V_i\|\ \|\varphi^{q+1}(x) - x^{\nu,*}\| \\ &\leq L_\nu\ \|V_i\|\ \rho\ \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}. \end{aligned}$$

Using the assumption on ε and the inequality $|V_i \cdot x^{\nu,*} - \eta_i| \geq 1$, we obtain

$$|V_i \cdot (\varphi^{q+1}(x) - x^{\nu,*})| \leq \rho\ |V_i \cdot x^{\nu,*} - \eta_i|$$

which implies $x \in D_\rho^{q+1}(x^{\nu,*})$. □

This theorem shows that the domains $D_\rho^q(x^{\nu,*})$ are extended domains of attraction of the network. It is easily proved that $D_\rho^p(x^{\nu,*})$ and $D_\rho^q(x^{\mu,*})$ are disjoint for any p and q if $\nu \neq \mu$.

The weights in the network are learned by minimizing the functional

$$J_i = \|V_i\|^2 \quad + \sum_{\nu \in I_{i,+}} \sum_{\mu \in I_{i,-}} (\eta_i^{\nu,\mu} - \eta_i)^2$$

$$+ \quad C\Big(\sum_{\nu \in I_{i,+}} (1 - V_i \cdot x^\nu + \eta_i)_+^2 + \sum_{\mu \in I_{i,-}} (1 + V_i \cdot x^\mu - \eta_i)_+^2 \Big) \quad (26)$$

for i such that $I_{i,+} = \{\nu \mid x_i^\nu = 1\} \neq \phi$ and $I_{i,-} = \{\nu \mid x_i^\nu = 0\} \neq \phi$.

V. Three-layer Autoassociative Memory Network

In the network (2), we choose $\ell = m$, where ℓ and m denote the numbers of output units and of stored patterns, respectively. And we assume that the i-output unit fires only when the i-stored pattern x^i is input in the network. That is,

$$f(W_i \cdot x^\nu - \theta_i) \begin{cases} \geq 1 - \varepsilon & \text{for} \quad \nu = i, \\ \leq \varepsilon & \text{for} \quad \nu \neq i. \end{cases} \quad (27)$$

Then we have $I_{i,+} = \{i\}$ and $I_{i,-} = \{\nu \mid \nu \neq i\}$. In the present case, Theorem 1 also holds and the regions $D_\rho(x^\nu)$ with $\ell = m$, $\nu = 1, 2, ..., m$, are mutually disjoint. Therefore, we can use such a network to solve classification problems. The condition (27) can be written as

$$V_i \cdot x^i - \eta_i \geq 1, \quad (28)$$

$$V_i \cdot x^\nu - \eta_i \leq -1 \quad \text{for} \quad \nu \neq i. \quad (29)$$

The weights of the network are obtained by minimizing the following functional:

$$J_i = \|V_i\|^2 + \sum_{\mu \neq i} (\eta_i^{i,\mu} - \eta_i)^2 + C\big((1 - V_i \cdot x^i + \eta_i)_+^2 + \sum_{\mu \neq i} (1 + V_i \cdot x^\mu - \eta_i)_+^2\big). \quad (30)$$

We construct here a three-layer autoassociative memory network by adding one more layer to the above specified neural network. From the added layer units, linear combinations of $\varphi_i(x)$ with new weights c_{ki} are output:

$$z_k = \sum_{i=1}^m c_{ki} \varphi_i(x) \equiv \psi_k(x), \qquad k = 1, 2, ..., n. \quad (31)$$

We put $\psi(x) = {}^t(\psi_1(x), \psi_2(x), ..., \psi_n(x))$ and $z = {}^t(z_1, z_2, ..., z_n)$. Then (31) may be written in the form

$$z = \psi(x)$$

which is a mapping from R^n into R^n. The weights c_{ki} are determined by the autoassociative condition

$$x^\nu = \psi(x^\nu), \qquad \nu = 1, 2, ..., m,$$

that is,

$$\sum_{i=1}^m \varphi_i(x^\nu) c_{ki} = x_k^\nu, \qquad \nu = 1, 2, ..., m.$$

The elements of the coefficient matrix $(\varphi_i(x^\nu))$ satisfy

$$1 - \varepsilon \leq \varphi_i(x^i) < 1 \qquad \text{for} \quad \nu = i$$

and

$$0 < \varphi_i(x^\nu) \leq \varepsilon \qquad \text{for} \quad \nu \neq i.$$

Therefore, the coefficient matrix almost equals to the unit matrix and we have

$$c_{ki} = x_k^i + O(\varepsilon).$$

Concerning the network $z = \psi(x)$, the following theorem holds.

Theorem 4: Suppose that V_i and η_i satisfy (28) and (29). Then we have for any x, $\widetilde{x} \in D_\rho(x^\nu)$,

$$|\psi_k(x) - \psi_k(\widetilde{x})| \leq \varepsilon^{1-\rho} \ln \frac{1}{\varepsilon} \sum_{i=1}^m |c_{ki}| \, |V_i \cdot (x - \widetilde{x})|, \qquad k = 1, 2, ..., m. \quad (32)$$

Especially, when $\widetilde{x} = x^\nu$, we have

$$|\psi_k(x) - \psi_k(x^\nu)| \leq \rho \, \varepsilon^{1-\rho} \ln \frac{1}{\varepsilon} \sum_{i=1}^m |c_{ki}| \, |V_i \cdot x^\nu - \eta_i|, \qquad k = 1, 2, ..., m. \quad (33)$$

Proof: The proof follows immediately from the definition of $\psi_k(x)$ and Theorem 1. □

In the same as in Theorem 3, we can derive extended domains of attraction for the above $D_\rho(x^\nu)$. In the present case, we do not need the contraction mapping theorem to show the existence of fixed points. Because the stored patterns x^ν themselves are fixed points of the function ψ.

Let ψ^0 be the identity mapping in R^n and define ψ^q for $q \geq 1$ by $\psi^q(x) = \psi(\psi^{q-1}(x))$ recursively. We put

$$E_\rho^q(x^\nu) = \{x \mid |V_i \cdot (\psi^q(x) - x^\nu)| \leq \rho |V_i \cdot x^\nu - \eta_i|, \quad i = 1, 2, ..., m \}.$$

We have one more theorem which corresponds to Theorem 3.

Theorem 5: Let us define τ by

$$\tau = \sqrt{\sum_{k=1}^n \Big(\sum_{i=1}^m |c_{ki}| \|V_i\|\Big)^2} \, \varepsilon^{1-\rho} \ln \frac{1}{\varepsilon}.$$

Then if ε is small enough so as to satisfy $\tau < 1$, the function ψ becomes a contraction mapping in $D_\rho(x^\nu)$.

Put

$$M_\nu = \sqrt{\sum_{k=1}^n \Big(\sum_{i=1}^m |c_{ki}| |V_i \cdot x^\nu - \eta_i|\Big)^2}.$$

Then we have for any $x \in E_\rho^q(x^\nu)$,

$$\|\psi^{q+1}(x) - x^\nu\| \le M_\nu \rho\, \varepsilon^{1-\rho}\, \ln\frac{1}{\varepsilon}. \tag{34}$$

Moreover, if ε is small enough so as to satisfy

$$M_\nu \max_{i=1,2,\ldots,m} \|V_i\|\, \varepsilon^{1-\rho}\, \ln\frac{1}{\varepsilon} \le 1,$$

we have the inclusions

$$D_\rho(x^\nu) = E_\rho^0(x^\nu) \subset E_\rho^1(x^\nu) \subset E_\rho^2(x^\nu) \subset \ldots \tag{35}$$

Proof: By estimating (32) in Theorem 4 from above and summing up both sides from $k = 1$ to n after squared, we obtain

$$\|\psi(x) - \psi(\widetilde{x})\| \le \tau\, \|x - \widetilde{x}\|.$$

This proves the first assertion.

The second and third assertions (34) and (35) are obtained in the same way as the proof in Theorem 3. □

By this theorem, we see that $E_\rho^q(x^\nu)$ are extended domains of attraction of the network. It is easily seen that $E_\rho^p(x^\nu)$ and $E_\rho^q(x^\mu)$ are disjoint for any p and q if $\nu \neq \mu$.

VI. Simulations

Using the two-layer and three-layer autoassociative memory networks constructed in Sections IV and V, respectively, we carry out simulations of character recognition.

First, we apply the two-layer network to the recognition of the alphabet. We store the following 26 alphabet in the network.

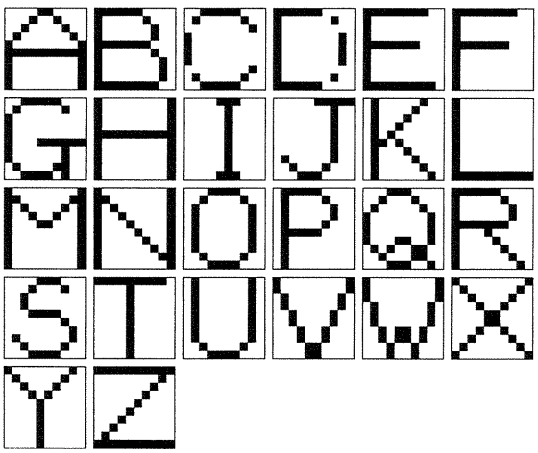

Fig. 1. Stored alphabet. Each character is represented by a 10 × 10 grid matrix. Black and white squares indicate 1 and 0, respectively.

The number of the input and output units is $n = 100$. Hence, we have 10100 connection weights. For the learning of the network, we minimize the functional (26) by the gradient method with step size $2.0 * 10^{-6}$. We select $\rho = 0.989$ and $\varepsilon = \exp(-1000)$. Then the value $\ln(1-\varepsilon)/\varepsilon$ equals to 1000 and the value of $\varepsilon^{1-\rho}\ln(1-\varepsilon)/\varepsilon$ is 0.0167.

We check the assumptions in Theorems 2 and 3. The contraction factor κ in Theorem 2 is 0.171 which is smaller than 1. Since $L_\nu \le 15.049$ for all ν and $\max_{i=1,2,\ldots,100} \|V_i\|\varepsilon^{1-\rho}\ln(1/\varepsilon) \le 0.026$, their product is smaller than 0.392, which satisfies the assumption of Theorem 3.

We shall try the recognition of the following noisy patterns:

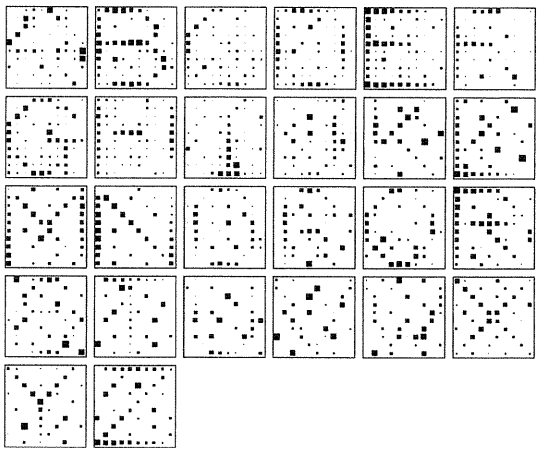

Fig. 2. 26 noisy patterns. Each pattern is represented by a 10 × 10 grid matrix. The area of a square is proportional to the gray level of the pattern.

Before inputting these patterns in the network, we normalize each pattern by dividing its maximum element. To check which domain of attraction they are contained in, we compute the ratio

$$r_{ineq}^2 = \max_{i=1,2,\ldots,100} \frac{|V_i \cdot (\varphi^q(x) - x^{\nu,*})|}{|V_i \cdot x^{\nu,*} - \eta_i|}$$

for $0 \le q \le 10$ and $\nu = 1, 2, \ldots, 26$, where we use x^ν in place of $x^{\nu,*}$. If $r_{ineq}^2 \le 0.989$ for some q and ν, the input pattern x is recognized as x^ν. In the table below, we list the results of recognition.

TABLE I

Results of Recognition. Top, Middle and Bottom in Each Box Represent the Values of q, r_{ineq}^2 and the Recognized Pattern. The Symbol × Denotes a Failure of Recognition.

4	10	2	3	5	3
0.767	1.685	0.748	0.000	0.792	0.775
A	×	C	D	E	F
10	3	3	1	2	3
1.367	0.766	0.825	0.886	0.889	0.000
×	H	I	J	K	L
1	2	3	3	3	10
0.578	0.476	0.000	0.785	0.870	1.295
M	N	O	P	Q	×
8	6	8	4	4	1
0.784	0.837	0.762	0.687	0.000	0.628
S	T	U	V	W	X
0	1				
0.907	0.787				
Y	Z				

We compare these results with those obtained by the two-layer autoassociative network which was learned under equality associative conditions. Since $|V_i \cdot x^\nu - \eta_i| = 1$ holds in this case, the ratio is given by

$$r_{eq}^2 = \max_{i=1,2,\ldots,100} |V_i \cdot (\varphi^q(x) - x^\nu)|.$$

If r_{eq}^2 is smaller than $\rho = 0.989$ for some q and ν, the input pattern x is recognized as x^ν. The following table shows the results of recognition.

TABLE II

RESULTS OF RECOGNITION. TOP, MIDDLE AND BOTTOM IN EACH BOX REPRESENT THE VALUES OF q, r_{eq}^2 AND THE RECOGNIZED PATTERN. THE SYMBOL × DENOTES A FAILURE OF RECOGNITION.

10 1.773 ×	4 0.063 B	10 1.945 ×	4 0.000 D	10 1.431 ×	10 2.100 ×
10 1.830 ×	10 2.117 ×	10 1.232 ×	10 1.531 ×	10 1.660 ×	10 1.894 ×
1 0.616 M	10 1.467 ×	10 1.054 ×	10 1.486 ×	10 1.655 ×	10 1.367 ×
10 1.714 ×	10 1.931 ×	10 1.930 ×	10 1.876 ×	10 1.898 ×	1 0.616 X
3 0.000 Y	10 1.547 ×				

As is easily seen from these results, the former is superior to the latter in the viewpoint of recognition ability.

Next, we apply the three-layer autoassociative memory network constructed in Section V to recognize the alphabet. The 26 alphabet in Figure 1 are stored to learn the network. It suffices to determine only 2626 weights which connect the input and the middle layers. The weights are determined by minimizing the functional (30). Let us choose $\varepsilon = \exp(-50)$ and $\rho = 0.800$. In this case, the value of $\varepsilon^{1-\rho}\ln(1-\varepsilon)/\varepsilon$ is 0.00227, and so τ in Theorem 5 is smaller than 0.145. Also $\max_{\nu=1,2,\ldots,26} M_\nu$ in Theorem 5 is 119.995, and hence $M_\nu \max_{i=1,2,\ldots,26} \|V_i\| \varepsilon^{1-\rho}\ln(1/\varepsilon) < 0.418$. Therefore, all the conditions in Theorem 5 are satisfied. To verify the recognition, the ratio

$$r_{ineq}^3 = \max_{i=1,2,\ldots,26} \frac{|V_i \cdot (\psi^q(x) - x^\nu)|}{|V_i \cdot x^\nu - \eta_i|}$$

is compared with the parameter $\rho = 0.800$. A simulation is carried out using the noisy patterns in Figure 2. In the simulation, the output $\varphi^q(x)$ at the middle layer for the noisy pattern x is normarized as $\varphi^q(x)/\max_{i=1,2,\ldots,26} \varphi_i^q(x)$ which is regarded as a new output vector at the middle layer. The table below shows the results.

TABLE III

RESULTS OF RECOGNITION. TOP, MIDDLE AND BOTTOM IN EACH BOX REPRESENT THE VALUES OF q, r_{ineq}^3 AND THE RECOGNIZED PATTERN. THE SYMBOL × DENOTES A FAILURE OF RECOGNITION.

1 0.000 A	1 0.000 B	1 0.034 ×	1 0.000 D	1 0.000 E	1 0.011 ×
1 0.001 G	1 0.022 ×	1 0.000 I	1 0.000 J	1 0.000 K	2 0.000 ×
1 0.000 M	1 0.000 N	1 0.000 O	1 0.000 P	1 0.000 Q	1 0.000 R
1 0.000 S	1 0.009 T	1 0.059 U	1 0.000 V	1 0.000 W	1 0.000 X
1 0.000 Y	1 0.000 Z				

These results are compared with those obtained by the three-layer autoassociative network which was learned under equality associative conditions. In this case, the ratio is given by

$$r_{eq}^3 = \max_{i=1,2,\ldots,26} |V_i \cdot (\psi^q(x) - x^\nu)|$$

by the same reason previously. The outputs at the middle layer are normarized as before. The table below shows the results of recognition.

TABLE IV

RESULTS OF RECOGNITION. TOP, MIDDLE AND BOTTOM IN EACH BOX REPRESENT THE VALUES OF q, r_{eq}^3 AND THE RECOGNIZED PATTERN. THE SYMBOL × DENOTES A FAILURE OF RECOGNITION.

1 0.000 A	1 0.000 B	1 0.061 ×	1 0.000 D	1 0.000 E	1 0.024 F
1 0.004 G	10 2.101 ×	1 0.000 I	1 0.020 ×	1 0.000 K	1 0.000 ×
1 0.590 M	1 0.000 N	1 0.000 O	1 0.000 P	1 0.000 Q	1 0.000 R
1 0.000 S	1 0.281 T	1 0.000 ×	1 0.004 ×	1 0.000 W	0 0.638 X
1 0.000 Y	1 0.379 Z				

Obviously, the former has higher recognition ability than the latter.

VII. Concluding Remarks

We proposed a learning method of associative memory networks. Our learning algorithm for the network is a minimizing process of a functional under inequality associative conditions for stored patterns. By relaxing equality associative conditions into inequality associative conditions, we could obtain the regions, each of which is mapped into a neighbor of an associative pattern, larger than those derived under equality associative conditions. These regions become domains of attraction in the case that the network is of autoassociative type. In this case, the network function can be shown to be a contraction mapping in the domains. The functional to be minimized was derived based on the shape of the obtained regions.

In the simulation, two kinds of autoassociative memory networks in this paper were applied to character recognition and their recognition ability was compared with that of the neural networks constructed under equality associative conditions in [8] and [9].

Our discussion is in a linear theory, because the network contains only one nonlinear layer. However, regions, each of which is mapped into a neighbor of an associative pattern, can be obtained under inequality associative conditions imposed on the final layer of a network, even if it is a multilayer network. These regions contain nonlinear functions and their shape is complicated. It is a future work to clarify such regions and to find learning algorithms for multilayer networks.

Appendix

Proof of Theorem 1: The proof is done by using essentially the same technique as in Theorem 1 in [8]. By the mean value theorem, we have

$$\varphi_i(x) - \varphi_i(\widetilde{x}) = f(z_i)(1 - f(z_i))W_i \cdot (x - \widetilde{x}), \qquad \text{(A1)}$$

where z_i is given by

$$z_i = \lambda(W_i \cdot x - \theta_i) + (1 - \lambda)(W_i \cdot \widetilde{x} - \theta_i), \qquad 0 < \lambda < 1.$$

We rewrite z_i as

$$\begin{aligned} z_i &= W_i \cdot x^\nu - \theta_i + W_i \cdot (\lambda x + (1-\lambda)\widetilde{x} - x^\nu) \\ &= (a_i^\nu + V_i \cdot (\lambda x + (1-\lambda)\widetilde{x} - x^\nu)) \ln \frac{1-\varepsilon}{\varepsilon}, \end{aligned}$$

where $a_i^\nu = V_i \cdot x^\nu - \eta_i$. We put

$$\begin{aligned} z_i^- &= (a_i^\nu - \rho\,|a_i^\nu|)\ \ln \frac{1-\varepsilon}{\varepsilon}, \\ z_i^+ &= (a_i^\nu + \rho\,|a_i^\nu|)\ \ln \frac{1-\varepsilon}{\varepsilon}. \end{aligned}$$

Since $\lambda x + (1-\lambda)\widetilde{x}$ belongs to $D_\rho(x^\nu)$, we have

$$z_i^- \le z_i \le z_i^+.$$

By the monotonicity of f, it holds that

$$f(z_i^-) \le f(z_i) \le f(z_i^+).$$

Therefore, we have for $i \in I_{i,-}$,

$$f(z_i) \le f((-1+\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon}) < \frac{1}{2}$$

and hence,

$$\begin{aligned} f(z_i)(1 - f(z_i)) &\le f((-1+\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon}) \\ &\quad \cdot(1 - f((-1+\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon})) \\ &= \frac{\exp(-(1-\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon})}{(1+\exp(-(1-\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon}))^2}. \end{aligned} \qquad \text{(A2)}$$

Since $|a_i^\nu| \ge 1$ and $\exp(-t)/(1+\exp(-t))^2$ is monotonically decreasing, we have

$$f(z_i)(1 - f(z_i)) \le \frac{\exp(-(1-\rho)\ \ln \frac{1-\varepsilon}{\varepsilon})}{(1+\exp(-(1-\rho)\ \ln \frac{1-\varepsilon}{\varepsilon}))^2}.$$

Using here the inequality $1/(1+\exp(-(1-\rho)\ln(1-\varepsilon)/\varepsilon))^2 \le (1-\varepsilon)^2$, the last term is bounded by $\varepsilon^{1-\rho}$. Therefore we obtain

$$f(z_i)(1 - f(z_i)) \le \varepsilon^{1-\rho}. \qquad \text{(A3)}$$

When $i \in I_{i,+}$, it follows that

$$\frac{1}{2} < f((1-\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon}) \le f(z_i).$$

Hence,

$$\begin{aligned} f(z_i)(1 - f(z_i)) &\le f((1-\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon}) \\ &\quad \cdot(1 - f((1-\rho)|a_i^\nu|\ \ln \frac{1-\varepsilon}{\varepsilon})). \end{aligned}$$

By an easy calculation, we see that the right hand side equals to the last term of (A2). Consequently, we obtain

$$f(z_i)(1 - f(z_i)) \le \varepsilon^{1-\rho}. \qquad \text{(A4)}$$

Combining (A3) and (A4) with (A1) gives the first result (8) of Theorem 1.

If we choose $\widetilde{x} = x^\nu$ in (8), we have for any $x \in D_\rho(x^\nu)$,

$$\begin{aligned} |\varphi_i(x) - \varphi_i(x^\nu)| &\le \varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon}\ |V_i \cdot (x - x^\nu)| \\ &\le \rho\varepsilon^{1-\rho}\ \ln \frac{1}{\varepsilon} |V_i \cdot x^\nu - \eta_i|, \end{aligned}$$

since x is in $D_\rho(x^\nu)$. This proves (9) of Theorem 1. □

References

[1] S.I. Amari, "Mathematical foundations of neurocomputing", *Proceedings of the IEEE*, vol. 78, pp.1443-1463, 1990.

[2] E.R. Caianiello and A. de Benedictis, "Neural associative memories with minimum connectivity", *Neural Networks*, vol. 5, pp.433-439, 1992.

[3] M. Cottrell, "Stability and attractivity in associative memory networks", *Biological Cybernetics*, vol. 58, pp.129-139, 1988.

[4] M.H. Hassoun, " Dynamic associative memories" in *Artificial Neural Networks and Statistical Pattern Recognition*, Amsterdam: North-Holland, 1991.

[5] M.H. Hassoun and J. Song, "Adaptive Ho-Kashyap rules for perceptron training", *IEEE Trans. Neural Networks*, vol. 3, pp.51-61, 1992.

[6] T. Kohonen, *Self-Organization and Associative Memory*, New York: Springer, 1984.

[7] R.J. McEliece, E.C. Posner, E.R. Rodemich, and S.S. Venkatesh, "The Capacity of the Hopfield Associative Memory," *IEEE Trans. Inform. Theory*, vol. 33, pp.461-482, 1987.

[8] K. Niijima, "Domains of attraction in autoassociative memory networks", *New Generation Computing*, vol. 12, pp.395-407, 1994.

[9] K. Niijima, "Domains of attraction in 3 layer autoassociative memory models for recognizing analog patterns", submitted to *Biological Cybernetics*.

[10] L.B. Rall, *Computational Solution of Nonlinear Operator Equations*, New York: Wiley, 1969.