KŌDAI MATH. SEM. REP.
19 (1967), 95–102

ON A LEARNING NETWORK

BY YASUICHI HORIBE

1. Introduction.

It has been recognized that the adaptive threshold (logic) elements or the threshold elements with variable parameters can be used as basic building blocks in pattern classifying learning system. The learning machine of simple perceptron type [2] which essentially consists of a single adaptive threshold element is an important example, having as input *stimuli or patterns,* finite binary vectors in full generality. The set of all possible stimuli is supposed to be pre-dichotomized into two classes for simplicity, and the state of this system is regarded as the set of variable parameters of the threshold element. Now by *learning* of the system we mean the change of the state based upon input stimuli presented to and their corresponding output responses of the system, so that the system may eventually correctly answer to which class the current stimulus belongs. Convergence of the state in finite number of steps to a desired state under a suitable learning algorithm is one of the principal results of the learning system of this type [1]. To assure this convergence, however, we necessarily have to put a rather strong assumption of linear separability of the stimulus classes.

In this paper we consider some aspects of a learning network with adaptive threshold elements and with a majority decision logic element. This system may be regarded as a generalized version of the above system with a single adaptive threshold element, to the effect that the former removes the condition of linear separability of stimulus classes and needs only that the stimulus classes are disjoint.

2. Formulation and difinitions.

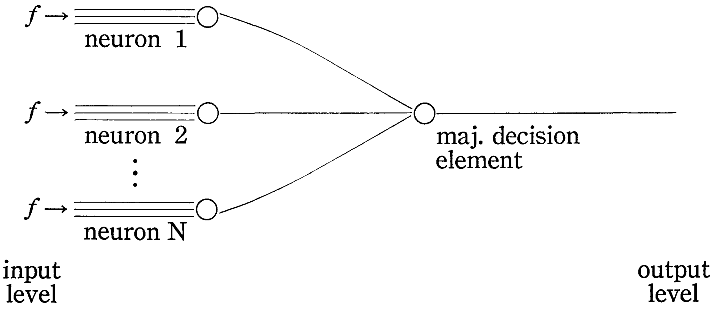
Consider a system (or an organ) which accepts as stimulus world, a subset F of the set Q^n of all binary sequences of length n, i.e. $Q^n=\underbrace{Q\times\cdots\times Q}_{(n)}$ for $Q=\{0, 1\}$.

The set F is pre-dichotomized into positive class F^+ and negative class F^- i.e. $F=F^+\cup F^-$ and $F^+\cap F^-=0$ (empty set).

The problem is to construct a system which can "learn" for any given stimulus $f\in F$ whether $f\in F^+$ or $f\in F^-$. (f will be considered a column vector from now on.) For this learning system we shall consider the following network using N (odd) adaptive threshold elements which are called *neurons* only for simplicity,

Received July 7, 1966.

and a single majority decision element. These elements are interconnected as in the figure, where all adaptive threshold elements are to accept stimulus f simultaneously.



Suppose that each neuron $j, j=1, 2, \cdots, N$, which has n-dimensional variable weight (column) vector w^j and variable threshold value θ^j behaves by the following rule:

$$\text{outputs} \begin{Bmatrix} \oplus \\ \ominus \end{Bmatrix} \text{ if } w^j(f) \begin{Bmatrix} > \\ \leqq \end{Bmatrix} \theta^j \text{ for input } f \in F^+,$$

$$\text{outputs} \begin{Bmatrix} \oplus \\ \ominus \end{Bmatrix} \text{ if } w^j(f) \begin{Bmatrix} \geqq \\ < \end{Bmatrix} \theta^j \text{ for input } f \in F^-$$

where $w^j(f)$ is the inner product of vectors w^j and f.

If we denote by $\pi(A)$, the number of elements in finite set A, then system output when stimulus f is presented to it is defined by the following majority decision rule:

$$\text{output } \oplus \text{ if } \pi\{j;\ \text{input } f \text{ causes neuron } j \text{ to output } \oplus\} \geqq \frac{N+1}{2},$$

$$\text{output } \ominus \text{ if } \pi\{j;\ \text{input } f \text{ causes neuron } j \text{ to output } \ominus\} \geqq \frac{N+1}{2}.$$

The system's output $\oplus$ ($\ominus$) for input f will be regarded as the system's decision that $f \in F^+$ ($f \in F^-$).

We have now that the system gives correct output response for $f \in F$ if and only if

$$\pi\{j;\ w^j(f) > \theta^j\} \geqq \frac{N+1}{2} \quad \text{for} \quad f \in F^+,$$

and

$$\pi\{j;\ w^j(f) < \theta^j\} \geqq \frac{N+1}{2} \quad \text{for} \quad f \in F^-.$$

The above statement may be simplified if we use the following "transformed" set of stimuli and the combination of weights and threshold value:

$$G^+=\left\{\begin{pmatrix} f^+ \\ -1 \end{pmatrix};\ f^+\epsilon F^+\right\},\qquad G^-=\left\{\begin{pmatrix} -f^- \\ 1 \end{pmatrix};\ f^-\epsilon F^-\right\},$$

$$G=G^+\cup G^-,\qquad \text{and}\qquad v^j=\begin{pmatrix} w^j \\ \theta^j \end{pmatrix},\quad j=1, 2, \cdots, N.$$

Then a mapping (one-to-one) φ: $F\rightarrow G$ is defined by

$$\varphi(f)=\begin{pmatrix} f \\ -1 \end{pmatrix}\epsilon G^+ \quad \text{for} \quad f\epsilon F^+,$$

and

$$\varphi(f)=\begin{pmatrix} -f \\ 1 \end{pmatrix}\epsilon G^- \quad \text{for} \quad f\epsilon F^-.$$

By this notation we see that the system gives correct output for $f\epsilon F$ if and only if $\pi\{j;\ v^j(g)>0\}\geqq(N+1)/2$ where $g=\varphi(f)$. The g of G will be also called stimulus.

Suppose that at each of the discrete time points $t=0, 1, 2, \cdots$ a stimulus g is presented to the system, taken from G in a certain manner, and stimulus at time t is denoted by g_t. The *system's state at time t* will be defined as the matrix

$$V_t=[v_t^1, v_t^2, \cdots, v_t^N].$$

Then the problem cited above at the beginning of the section reads as follows: by a suitable learning (or training) algorithm which properly changes the state of the system, whether it is possible to lead the system to a state V_{t_0} for some t_0 such that at any time satisfying $t\geqq t_0$ we have $\pi(I_t)\geqq(N+1)/2$, where $I_t=\{j;\ v_t^j(g_t)>0\}$.

3. Consistency.

Let us call *classification*, a pair of sets (F^+, F^-) such that $F^+\subset Q^n$, $F^-\subset Q^n$, and $F^+\cap F^-=0$, and suppose that a certain classification $C=(F^+, F^-)$ forms the set of all stimuli acceptable by the system.

The *linearly separable classification* is a classification such that there exists a vector v satisfying $v(\varphi(f))>0$ for any $f\epsilon F^+\cup F^-$. Geometrically, this condition is equivalent to the existence of a hyperplane strictly separating two sets F^+ and F^-.

It is obviously necessary, when considering the proposed problem, that the following statement (C) holds for the given classification $C=(F^+, F^-)$, since, if not, we can not find any state where the system gives correct response for each stimulus.

(C) *There exists a state of the system* $\tilde{V}=[\tilde{v}^1, \cdots, \tilde{v}^N]$ *such that* $\pi(\tilde{I}_g)\geqq(N+1)/2$ *for any* $g\epsilon G$ *where* $\tilde{I}_g=\{j;\ \tilde{v}^j(g)>0\}$.

The next theorem shows, however, that the system is "universal" for any classification, if it can have suitable number of neurons.

THEOREM 1. *For two arbitrary subsets F^+, F^- of Q^n there exists a state $\tilde{V}$ satisfying* (C) *described above with suitably chosen number N of neurons, if and only if $F^+ \cap F^- = 0$.*

Proof. Necessity.

Suppose that F^+ and F^- are not disjoint. Put $g=\begin{bmatrix} f \\ -1 \end{bmatrix}$ for some $f \in F^+ \cup F^-$, then $g \in G$ and $-g \in G$, and also $\pi(\tilde{I}_g) \geqq (N+1)/2$ and $\pi(\tilde{I}_{-g}) \geqq (N+1)/2$. Since

$$\tilde{I}_{-g}=\{j;\ \tilde{v}^j(-g)>0\}=\{j;\ \tilde{v}^j(g)<0\}\subset I-\tilde{I}_g,$$

where $I=\{1, 2, \cdots, N\}$, we have

$$\pi(\tilde{I}_{-g}) \leqq N-\pi(\tilde{I}_g) \leqq N-\frac{N+1}{2}=\frac{N+1}{2}-1<\frac{N+1}{2},$$

a contradiction.

Sufficiency.

Suppose that $F^+ \cap F^- = 0$. If F^+ and F^- are linearly separable, then it is clear that there exists a state $\tilde{V}$ such that $\pi(\tilde{I}_g)=N \geqq (N+1)/2$ for any $g \in G$.

Remark here that for any $f \in Q^n$, classification $(Q^n-\{f\}, \{f\})$ is linearly separable. We therefore see that the classification $(F^+ \cup F^- - \{f\}, \{f\})$, $f \in F^-$, for the given general not necessarily linearly separable classification (F^+, F^-) is linearly separable, hence has a state satisfying (C). Using mathematical induction, the only thing to be proved is that for an arbitrary classification $C=(F^+, F^-)$ having a state (with N neurons) satisfying (C), if we choose suitable number of neurons, we can have a state satisfying (C) for the classification $C'=(F^+ \cup \{f\}, F^- - \{f\})$, $f \in F^-$.

Now suppose that for the classification C, there exists a state $\tilde{V}=[\tilde{v}^1, \cdots, \tilde{v}^N]$ such that $\pi(\tilde{I}_g) \geqq (N+1)/2$ for any $g \in G$. Putting $\begin{bmatrix} f \\ -1 \end{bmatrix}=g$ as usual, then $-g \in G^-$, since $f \in F^-$. We have that $\varphi(F^+ \cup F^-)=(G^+ \cup \{g\}) \cup (G^- - \{-g\})$.

Assume for this g, $\pi(\tilde{I}_g)=(N+1)/2-s$. If $s=0$, then it is clear that the state $\tilde{V}$ satisfies (C) for the classification C', hence s assumes $1, 2, \cdots, (N+1)/2$.

Now add $2s$ number of neurons newly to the system with N number of neurons having the state $\tilde{V}$ satisfying (C). We shall show that there is a state $\hat{\tilde{V}}=[\bar{v}^1, \cdots, \bar{v}^{N+2s}]$ satisfying (C) for the classification C'. Since there exists a hyperplane which has the set $F^+ \cup \{f\}$ in one of the half spaces determined by it, we can put

$$\bar{v}^{N+1}=\bar{v}^{N+2}=\cdots=\bar{v}^{N+s}=v_1$$

such that for any $h \in G^+ \cup \{g\}$ we have $v_1(h)>0$.

Since also there exists a hyperplane which separates the two sets $F^- - \{f\}$ and $\{f\}$, we can put

$$\bar{v}^{N+s+1}=\bar{v}^{N+s+2}=\cdots=\bar{v}^{N+2s}=v_2$$

such that for any $h\in G^- -\{-g\}$ we have $v_2(h)>0$, and $v_2(g)>0$.

Finally put

$$\bar{v}^1=\tilde{v}^1,\ \bar{v}^2=\tilde{v}^2,\ \cdots,\ \bar{v}^N=\tilde{v}^N.$$

Then the following relations are immediate consequences of the above argument:

$$\pi(\bar{I}_h)\geqq \frac{N+1}{2}+s=\frac{N'+1}{2} \quad \text{for} \quad h\in G^+,$$

$$\pi(\bar{I}_g)\geqq\left(\frac{N+1}{2}-s\right)+s+s=\frac{N'+1}{2} \quad \text{for} \quad g,$$

$$\pi(\bar{I}_h)\geqq \frac{N+1}{2}+s=\frac{N'+1}{2} \quad \text{for} \quad h\in G^- -\{-g\},$$

where $N'=N+2s$, $\bar{I}_h=\{j;\ \bar{v}^j(h)>0\}$. q.e.d.

4. On number of neurons.

The proof of the theorem 1 gives a suggestion for an enumeration of number of neurons that are necessary for assuring the existence of a state which satisfies (C). For instance, if we know only $\pi(F^+)$ and $\pi(F^-)$, a rough upper bound of neuron number which may guarantee the existence of a state satisfying (C), for any classification (F^+, F^-) with given values $\pi(F^+)$ and $\pi(F^-)$, will be readily calculated as follows:

Assume that $\pi(F^+)\geqq\pi(F^-)=q$. Denote a sufficient number of neurons by N_r for the classification

$$C_r=(F^+\cup(F^- -\{f_1, \cdots, f_r\}),\ \{f_1, \cdots, f_r\})$$

where $\{f_1, \cdots, f_r\}\subset F^-$.

It is clear that $N_1=1$, since C_1 is linearly separable. If we take maximum $(N+1)/2$ for s in the proof of the theorem 1, then the following relation is readily seen:

$$N_{r+1}=N_r+2\cdot\frac{N_r+1}{2}.$$

Therefore we have $N_r=2^r-1$, and when $r=\pi(F^-)$ we can reach our first classification (F^+, F^-), hence 2^q-1.

THEOREM 2. *For any classification* (F^+, F^-) *with fixed values* $\pi(F^+)$ *and* $\pi(F^-)$, *we have as a sufficient number of neurons which assures the existence of a state satisfying* (C), *the number* 2^q-1 *at most, where* $q=\min(\pi(F^+), \pi(F^-))$.

A more efficient way for enumeration of neuron number may be as follows. Form the set C of all classifications (E^+, E^-) each of which satisfies that $E^+ \subset F^+$, $E^- \subset F^-$, and such that

$$((F^+ - E^+) \cup E^-, (F^- - E^-) \cup E^+)$$

is linearly separable. If we put

$$\min_{(E^+, E^-) \in C} (\pi(E^+) + \pi(E^-)) = q_0,$$

then $2^{q_0} - 1$ neurons are sufficient for (F^+, F^-).

5. A convergence theorem.

Suppose now that a system is given such that it has as stimulus world, a classification (F^+, F^-), and has the existence of a state $\tilde{V} = [\tilde{v}^1, \cdots, \tilde{v}^N]$ (with N neurons) satisfying (C). Hence supposing that

$$\varphi(F^+) = G^+ = \{g_1, \cdots, g_k\}, \quad \varphi(F^-) = G^- = \{g_{k+1}, \cdots, g_{k+l}\},$$

(Notations g_i and g_t may not confusing.) we have $\pi(\tilde{I}_i) \geqq (N+1)/2$ for any $i = 1, \cdots, k+l$, where

$$\tilde{I}_i = \{j;\ \tilde{v}^j(g_i) > 0\}.$$

Note that if $\cap_{i=1}^{k+l} \tilde{I}_i \neq 0$, then there exists a neuron j such that $\tilde{v}^j(g_i) > 0$ for any $i = 1, \cdots, k+l$, hence (F^+, F^-) is linearly separable for which a single neuron $(N=1)$ is sufficient.

It is obvious that the sequence of sets $\{\tilde{I}_i\}$, $i = 1, \cdots, k+l$, such that there is a state $\tilde{V}$ satisfying (C), is not unique, for the given neuron number N and the given classification (F^+, F^-). Observe that the order of the arrangement of neurons is a trifle.

If we know one such sequence of sets $\{\tilde{I}_i\}$, $i = 1, \cdots, k+l$, then the following learning algorithm (A) may be adopted for the system.

(A) *Let the initial state i.e. the state at time* $t=0$ *be*

$$V_0 = [0, \cdots, 0].$$

The system in state $V_t = [v_t^1, \cdots, v_t^N]$ *at time* t *is presented stimulus* $g_t \in G$.

If $\pi(I_t) \geqq (N+1)/2$, *set* $V_{t+1} = V_t$ *for the state at time* $t+1$.

If $\pi(I_t) < (N+1)/2$, *then choose an arbitrary set* J_t *of* $(N+1)/2 - \pi(I_t)$ *neurons out of the set* $I - I_t$ *such that* $J_t \subset \tilde{I}_{g_i} = \{j;\ \tilde{v}^j(g_t) > 0\}$, *and "correct" the neurons in* J_t *as*:

$$v_{t+1}^j = v_t^j + g_t \quad \textit{for} \quad j \in J_t.$$

THEOREM 3. *When a sequence of sets* $\{\tilde{I}_i\}$, $i=1,\cdots,k+l$, *such that there is a state* $\tilde{V}$ *satisfying* (C) *is known, then by the learning algorithm* (A) *the system "learns" the classification* (F^+, F^-) *with finite number of corrections. In this case the stimulus sequence* $\{g_t\}$ *may be arbitrary.*

Proof. It is sufficient to assume that the sequence of times $t=0,1,2,\cdots$ (which may be finite or infinite) are those times when a correction is performed to the system, i.e. at least one neuron in the system is corrected.

Suppose that this sequence of times is infinite. Then we have a infinite sequence of sets $\{J_t\}$, $t=0,1,\cdots$. Now since $\cup_{t=0}^{\infty}J_t\subset I$ and I is finite set, there must exist some neuron j and a infinite subsequence $\{t_\nu\}$ of $\{t\}$ such that at any time in $\{t\}-\{t_\nu\}$ the neuron j is never corrected and $j\in J_{t_\nu}$, $\nu=1,2,\cdots$. Therefore we have $v^j_{t_{\nu+1}}=v^j_{t_\nu}+g_{t_\nu}$, $\nu=1,2,\cdots$. Put $t_{\nu+1}=\nu+1$ for simplicity, then

$$v^j_{\nu+1}=v^j_\nu+g_\nu. \tag{1}$$

Now the excellent proof of Novikoff for the simple perceptron convergence theorem [3] will be performed to complete our proof more succinctly.

If $\tilde{v}^j$ is multiplied to the both sides of (1), we have

$$v^j_{\nu+1}(\tilde{v}^j)=v^j_\nu(\tilde{v}^j)+\tilde{v}^j(g_\nu). \tag{2}$$

Since $J_t\subset\tilde{I}_{g_t}$ for any t, we have $\tilde{v}^j(g_\nu)>0$. If we put $\min_{g_\nu\in\{g_\nu\}}\tilde{v}^j(g_\nu)=m$, then $m>0$. It follows from (2) that

$$v^j_{\nu+1}(\tilde{v}^j)\geqq v^j_\nu(\tilde{v}^j)+m,$$

and hence

$$v^j_{\nu+1}(\tilde{v}^j)\geqq\nu m, \text{ since } v^j_0=0. \tag{3}$$

On the other hand, from (1) we have

$$||v^j_{\nu+1}||^2=||v^j_\nu||^2+2v^j_\nu(g_\nu)+||g_\nu||^2. \tag{4}$$

Since the neuron j is corrected at time ν, we have $v^j_\nu(g_\nu)\leqq0$. If we put $\max_{g\in G}||g||^2=M^2$,

then $$(|v^j_{\nu+1}||^2\leqq||v^j_\nu||^2+M^2,$$

hence

$$||v^j_{\nu+1}||^2\leqq\nu M^2, \text{ since } v^j_0=0. \tag{5}$$

From (3), (5), and the inequality

$$v^j_{\nu+1}(\tilde{v}^j)\leqq||\tilde{v}^j||\cdot||v^j_{\nu+1}||,$$

we have, putting $||\tilde{v}^j||=c$,

$$\nu m=cM\sqrt{\nu}\,.$$

which does not hold for large ν, a contradiction. q.e.d.

COROLLARY. *In the theorem* 3, *if each stimulus in* G *is to be presented to the system infinitely many times in the stimulus sequence* $\{g_t\}$, *then in finite number of stimulus presentations the system can reach a state which satisfies* (C).

The author is indebted to Professor K. Kunisawa for his encouragements and suggestions.

REFERENCES

[1] BLOCK, H. D., The perceptron, a model for brain functioning. Reviews of Modern Physics **34** (1962), 123-135.

[2] HORIBE, Y., On an adaptive process for learning finite patterns. Kōdai Math. Sem. Rep. **19** (1967), 43-52.

[3] NOVIKOFF, A. B. J., On convergence proofs for perceptrons. Proc. of the Symp. on Math. Theory of Automata (1962), 615-622, Polytechnic Press.

DEPARTMENT OF MATHEMATICS,
TOKYO INSTITUTE OF TECHNOLOGY.