A geometric learning algorithm for elementary perceptron and its convergence analysis

著者	Miyoshi Seiji, Nakayama Kenji
journal or publication title	IEEE International Conference on Neural Networks - Conference Proceedings
volume	3
page range	1913-1918
year	1997-06-01
URL	http://hdl.handle.net/2297/6848

A Geometric Learning Algorithm for Elementary Perceptron and Its Convergence Analysis

Seiji MIYOSHI*† and Kenji NAKAYAMA‡
*Graduate School of Natural Science and Technology, Kanazawa Univ.
†Department of Electronic Engineering, Kobe City College of Technology
‡Department of Electrical and Computer Eng. Faculty of Eng., Kanazawa Univ.
*‡2–40–20, Kodatsuno, Kanazawa 920, Japan
†8–3, Gakuen-Higashimachi, Nishi-ku, Kobe 651–21, Japan
*E–mail : miyoshi@kobe-kosen.ac.jp

ABSTRACT

In this paper, the geometric learning algorithm (GLA) is proposed for an elementary perceptron which includes a single output neuron. The GLA is a modified version of the affine projection algorithm (APA) for adaptive filters. The weights update vector is determined geometrically towards the intersection of the k hyperplanes which are perpendicular to patterns to be classified. k is the order of the GLA. In the case of the APA, the target of the coefficients update is a single point which corresponds to the best identification of the unknown system. On the other hand, in the case of the GLA, the target of the weights update is an area, in which all the given patterns are classified correctly. Thus, their convergence conditions are different. In this paper, the convergence condition of the 1st order GLA for 2 patterns is theoretically derived. The new concept "the angle of the solution area" is introduced. The computer simulation results support that this new concept is a good estimation of the convergence properties.

1. Introduction

The perceptron learning is well known as the learning algorithm for an elementary perceptron. This algorithm has a special merit, that is, if the given pattern set is linearly separable, the learning always finds the solution in a finite learning steps. This property is well known as the perceptron convergence theorem[1]. However, this algorithm has some demerits, that is, the learning is slow and the solution is not always excellent for noisy pattern classification.

On the other hand, the affine projection algorithm (APA) is well known as the generalized algorithm of the normalized LMS algorithm into the block signal processing in the field of adaptive filters[2]. The 2nd order APA was applied to the bidirectional associative memory (BAM) learning. The learning is faster than the perceptron learning[3]. However, the paper[3] is limited to the slightly special case, that is the BAM, and the conditions for the learning convergence have not been investigated in details.

In this paper, the geometric learning algorithm (GLA) is proposed for an elementary perceptron. The GLA is a modified version of the APA, that is, the GLA is applied to an elementary perceptron. The condition for the convergence within a finite number of learning steps of the 1st order GLA for 2 patterns is derived theoretically. After that, the new concept "the angle of the solution area" is introduced in order to estimate the convergence property of the given pattern set with many patterns. The goodness of this new concept is investigated through computer simulation. In this paper, the word "convergence" means that the learning process finishes by reaching the solution area.

2. Elementary perceptron

Figure 1 shows an elementary perceptron proposed by Rosenblatt[4]. The operation can be described by Eqs.(1) and (2).

$$u = \sum_{i=0}^{N-1} w_i x_i \qquad (1)$$

$$y = \begin{cases} +1, & u \geq 0 \\ -1, & u < 0 \end{cases} \qquad (2)$$

When 0 is substituted in u of Eq.(1), this means the hyperplane of which gradient and position are determined by the connection weights w_i. Therefore, an elementary perceptron has the ability to discriminate

0-7803-4122-8/97 $10.00©1997 IEEE

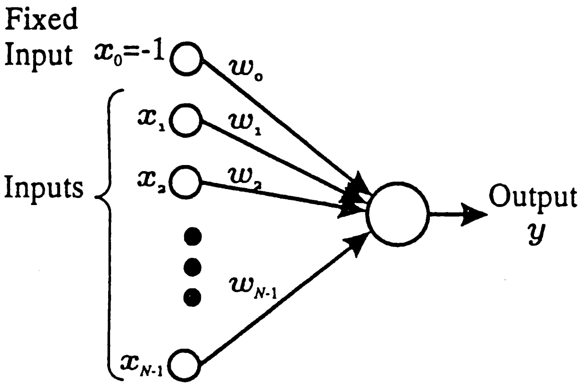

Fig. 1: Elementary perceptron.

the two classes which are divided by this hyperplane in the pattern space. In Fig.1, the threshold is fixed to 0 by equipping with the input x_0 which always takes -1 and the weight w_0. This type is convenient because the hyperplane includes the origin point in exchange for only 1 dimensional enlargement of input vectors. Therefore, this type of elementary perceptron is considered in this paper.

3. Geometric Learning Algorithm

3.1. LMS and Normalized LMS algorithms

The LMS algorithm is well known in the field of the adaptive filters. In this algorithm, the filter coefficients vector $h(n)$ is updated by[5]

$$h(n+1) = h(n) + \mu e(n)u(n) \qquad (3)$$

n is time step, $e(n)$ is the error, $u(n)$ is the tap input vector and μ is the step-size parameter. The LMS algorithm doesn't need to measure the correlation function nor to calculate the inverse matrix.

The normalized LMS (NLMS) algorithm was proposed independently by Nagumo and Noda[6], Albert and Gardner[7]. The filter coefficient vector $h(n)$ is updated by

$$h(n+1) = h(n) + \frac{\alpha}{\|u(n)\|^2}e(n)u(n) \qquad (4)$$

The NLMS algorithm is convergent in the mean-square sense if and only if the adaptation constant α satisfies[8]

$$0 < \alpha < 2 \qquad (5)$$

The NLMS algorithm is faster than the LMS algorithm[9]. The convergence rate is independent of the input signal.

3.2. Affine Projection Algorithm (APA)

The APA was proposed for an algorithm of adaptive filters[2]. Figure 2 shows the 2nd order APA conceptually. h^* is the target of the adaptation, that is the best identification of the unknown system. The update vector of the filter coefficients is taken to be perpendicular towards the intersection of the hyperplanes. The hyperplanes are the sets of the coefficients which correspond to the desired outputs at the different time. Therefore, the hyperplanes are perpendicular to the tap input vectors u^1,u^2 of the adaptive filter at each time. In the kth order APA (k–APA), k hyperplanes are used in each update. Figure 2 shows the case that the order is 2 and the dimension of the pattern vectors u^1,u^2 is 3. Therefore, the intersection of the hyperplanes is 1-dimensional line l. The ratio $\overline{PQ}/\overline{PO}$ is constant in the APA. The NLMS algorithm is equivalent to 1–APA. The APA converges if and only if the ratio satisfies[2],

$$0 < \frac{\overline{PQ}}{\overline{PO}} < 2 \qquad (6)$$

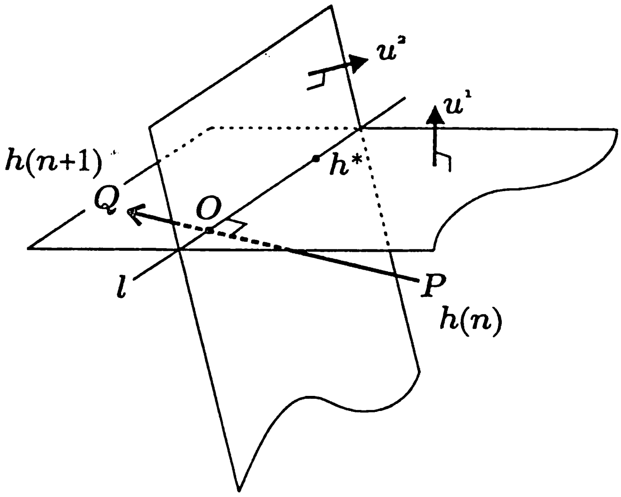

Fig. 2: 2nd-order APA.

3.3. Geometric Learning Algorithm (GLA)

As described in Sec.2, the weight hyperplane of an elementary perceptron shown in Fig.1 includes the origin point. Therefore, the intersection of such hyperplanes also includes the origin point. In this case, the intersection is "subspace", and doesn't have to be said "affine subspace".

For this reason, the algorithm, in which the APA is applied to the learning of an elementary perceptron for pattern classification is newly called "the geometric learning algorithm (GLA)". This name originates from the fact that the ratio of the update vector($\overline{PQ}$) and the perpendicular segment from the weight vector to the intersection($\overline{PO}$) is constant, that is "geometric". In this paper, the ratio $\overline{PQ}/\overline{PO}$ is called the learning constant and is denoted λ. Application of the k–APA is called the k–GLA. Application of k–APA means that the weights update is done by using k patterns which need more learning, that is the patterns of which classes don't agree with Eqs.(1) and (2). The k–GLA is described as follows:

```
begin
  $w(0)$ is randomly set;
  while $k_0 (> 0)$ patterns, of which classes don't
  agree with Eqs.(1) and (2), remain do begin
    if $k_0 \geq k$
          $X = (x^1, x^2, \cdots, x^k)^T$          (7)
    else
          $X = (x^1, x^2, \cdots, x^{k_0})^T$          (8)
    end if;
          $w(n+1) = w(n) - \lambda X^+ X w(n)$          (9)
  end while;
end;
```

X^+ means the Moore–Penrose generalized inverse of X. In Eqs.(7) and(8), $x^1 \sim x^k$ or $x^1 \sim x^{k_0}$ are selected from the patterns of which classes don't agree with Eqs.(1) and (2). It isn't defined here how to select the patterns.

For example, the weight vector is updated as follows in the 1–GLA:

$$w(n+1) = w(n) - \lambda \frac{x^T w(n)}{\|x\|^2} x \tag{10}$$

x is the pattern vector selected from the patterns of which classes don't agree with Eqs.(1) and (2).

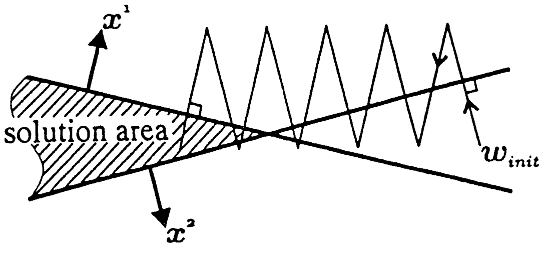

Fig. 3: Weights update process in perceptron learning.

Since the APA is the algorithm of adaptive filters, the target of the adaptation is a "single point". On the contrary, since the GLA is applied to an elementary perceptron, the target of the learning is an "area" in which the elementary perceptron classifies all the patterns correctly. Therefore, the convergence condition of the GLA is different from that of the APA given by Eq.(6). The analysis of the condition for the GLA convergence is the main subject of this paper and is described in the next section.

Figure 3 shows the perceptron learning. Figure 4 shows the 1–GLA. The circle in Fig.4 means that the norm of the weight vector is normalized at each learning step. Since the threshold is zero, the generality isn't lost.

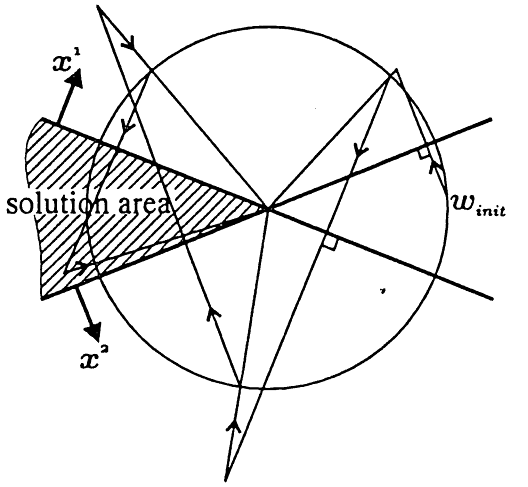

Fig. 4: Weights update process in 1-GLA.

4. Theoretical analysis of 1–GLA convergence

In this section, the convergence condition of the 1–GLA is derived. The 1–GLA is the application of the 1–APA to an elementary perceptron. This is the starting point of the theoretical analysis of the k–GLA.

As described in Sec.1, it had been proved that the perceptron learning always converges within a finite number of steps if the given patterns are linearly separable. On the contrary, the GLA doesn't always converge within a finite number of steps. Figure 5 shows the situation, in which the learning process oscillates, that is, ($\rightarrow A \rightarrow B \rightarrow C \rightarrow D \rightarrow$). If the learning falls into the oscillation, the weight cannot approach the solution area any more, that is, the learning doesn't converge. In the following, the condition that the oscillation occurs is analyzed when 2 patterns' classification is learned by the 1–GLA as the most basic case. The condition that the learning converges in a finite number of steps regardless of the initial weights is derived. Assuming that Fig.5 shows the 2–dimensional plane which includes the origin point and the 2 patterns to be classified in the $N(\geq 2)$–dimensional space, the arguments in this section are true for the $N(\geq 2)$ dimensional patterns. When $\lambda \leq 1$, the weight cannot cross the pattern hyperplane, that is, the weight cannot reach the solution area. Therefore, the case that $\lambda > 1$ is considered in the following.

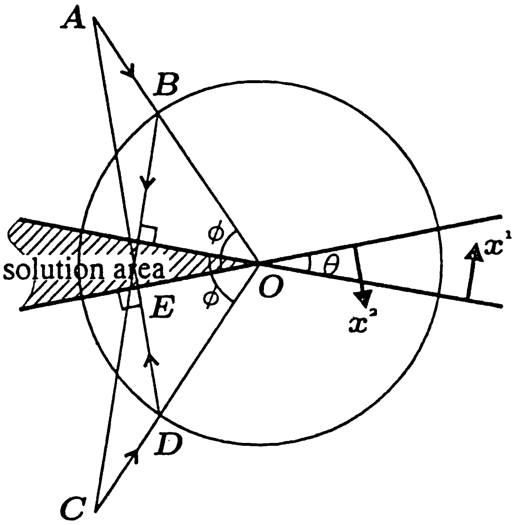

Fig. 5: Oscillation phenomenon in weights update process.

In Fig.5, $\overline{DA}/\overline{DE}$ is λ. As the two right–angled triangles $\triangle OED$, $\triangle OEA$ have the common side OE, the following condition is obtained.

$$(\lambda - 1)\sin(\phi - \theta)\cot\phi = \sin(\phi - \theta)\cot(\phi - \theta) \quad (11)$$

By solving this equation for λ, we get

$$\lambda = \frac{\tan\phi}{\tan(\phi - \theta)} + 1 \quad (12)$$

θ is determined by 2 patterns to be classified and their classes. If there exist λ and $\phi(> \theta)$ satisfying Eq.(12), the oscillation may occur.

Let consider the right side of Eq.(12) as the function f of ϕ.

$$f(\phi) = \frac{\tan\phi}{\tan(\phi - \theta)} + 1 \quad (13)$$

Figure 6 shows the outline of the function $f(\phi)$ when $0 < \theta \leq \frac{\pi}{2}$.

From Fig.6, in the case of $\lambda \geq \frac{\tan(\frac{\pi}{4}+\frac{\theta}{2})}{\tan(\frac{\pi}{4}-\frac{\theta}{2})} + 1$ or $\lambda \leq \frac{\tan(\frac{\pi}{4}-\frac{\theta}{2})}{\tan(\frac{\pi}{4}+\frac{\theta}{2})} + 1$, there exists two values (one value if equal sign) of $\phi(\geq \theta)$ satisfying $f(\phi) = \lambda$ when $0 < \theta \leq \frac{\pi}{2}$. This means two (one) oscillation states exist.

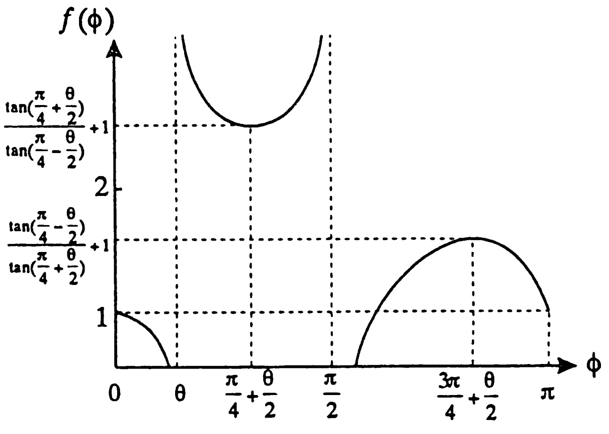

Fig. 6: Function f of ϕ $(0 < \theta \leq \frac{\pi}{2})$.

. Though the details are omitted here, it is proved that one of the two oscillations is stable and the other is unstable (stable only from one side if equal sign). Therefore, when $\lambda \geq \frac{\tan(\frac{\pi}{4}+\frac{\theta}{2})}{\tan(\frac{\pi}{4}-\frac{\theta}{2})} + 1$ or $\lambda \leq \frac{\tan(\frac{\pi}{4}-\frac{\theta}{2})}{\tan(\frac{\pi}{4}+\frac{\theta}{2})} + 1$, it is possible that the learning falls into the stable oscillation before reaching the solution area. Then, whether the learning converges or not depends on the initial weights.

Figure 6 shows that if $\frac{\tan(\frac{\pi}{4}+\frac{\theta}{2})}{\tan(\frac{\pi}{4}-\frac{\theta}{2})} + 1 > \lambda > \frac{\tan(\frac{\pi}{4}-\frac{\theta}{2})}{\tan(\frac{\pi}{4}+\frac{\theta}{2})} + 1$, $\phi(\geq \theta)$ satisfying $f(\phi) = \lambda$ doesn't exist. Though the details are also omitted here, investigating the direction of learning, it is proved that the learning always converges within a finite number of steps in this case.

Though the details are also omitted here, it is proved that the learning always converges within a finite number of steps when $\frac{\pi}{2} < \theta \leq \pi$.

From the above consideration, the condition that the learning converges regardless of the initial weights is given as follows:

- If $0 < \theta \leq \frac{\pi}{2}$,

$$\frac{\tan(\frac{\pi}{4}+\frac{\theta}{2})}{\tan(\frac{\pi}{4}-\frac{\theta}{2})} + 1 > \lambda > \frac{\tan(\frac{\pi}{4}-\frac{\theta}{2})}{\tan(\frac{\pi}{4}+\frac{\theta}{2})} + 1 \quad (14)$$

- If $\frac{\pi}{2} < \theta \leq \pi$,

$$\lambda > 1 \quad (15)$$

5. Convergence property of 1–GLA for many patterns

5.1. Angle of solution area

In Sec.4, it is proved that the convergence property of the 1–GLA for 2 patterns is determined by θ in Fig.5. That is, the larger θ is, the wider the range of λ for convergence is, as shown in Eqs.(14) and (15).

Comparing the case of more than 2 patterns (expressed as "many patterns" in this paper) with the case of 2 patterns, the learning process until convergence or oscillation is more complicated. The reasons are that the shape of the solution area is more complicated and the order of patterns to be presented is added to the degrees of freedom.

However, even in the case of many patterns, the convergence property is considered to be determined by a certain angle which is unique for the patterns to be classified. It is a future subject what angle corresponds to θ and how the angle is calculated. In this paper, the new concept "the angle of the solution areas" is introduced. This angle is relatively easy to calculate from the..patterns to be classified. The new concept includes θ in the case of 2 patterns.

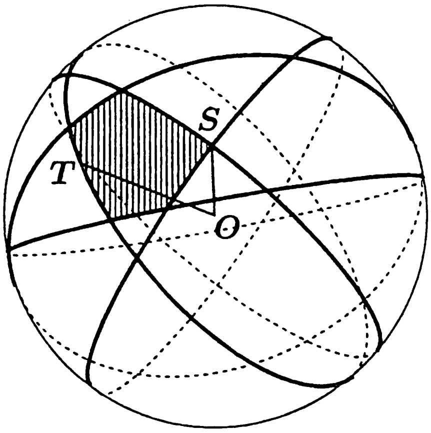

Fig. 7: Angle of solution area.

Figure 7 shows the solution area which is determined by 5 patterns. In this figure, the shaded area means the solution area. In Fig.7, S and T move on the circumference of the area. Let $\psi(S)$ be the maximum value of $\angle SOT$ when T circulates around the area with the fixed S. Then, the angle of the solution area ψ_{min} is defined to be the minimum value of $\psi(S)$ when S circulates around the area. ψ_{min} can be considered to be the minimum angle of the solution area viewed from the origin point.

The way of thinking to calculate the angle of the solution area ψ_{min} about the patterns to be classified is described as follows: The hyperplane being perpendicular to the weight vector separates the patterns correctly if and only if the weight vector is in the solution area. Figure 8 shows the relation between the patterns and the separating hyperplane. In this figure, the white patterns and the black patterns belong to the different classes. Therefore, the hyperplane separates these two classes correctly if and only if the weight vector is in the solution area. ψ_1 is the angle of one black pattern and the wedge–shaped range between two white patterns. ψ_2 is the angle of one reversed white pattern and the wedge-shaped range between two white patterns. As the angle of the solution area ψ_{min} is the minimum angle of the solution area viewed from the origin point, ψ_{min} can be calculated as the minimum angle of the ψ_1 and ψ_2 of all the pattern combinations. All the pattern combinations include the case of black and white interchange. In N–dimensional space, the wedge–shaped range of linear combinations with non–negative coefficients of 1 pattern, 2 patterns, 3 patterns, $\cdots$, $N-1$ patterns must be considered.

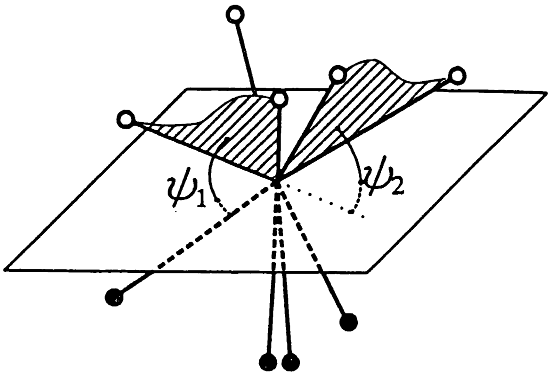

Fig. 8: Relation between patterns and separating hyperplane.

5.2. Computer simulation

The goodness of the angle of the solution area ψ_{min} as the estimation of the convergence properties is investigated through computer simulation.

First, 10 pattern sets are generated. Each set is composed of 5 patterns which are 7–dimensional. About all patterns, the 1st elements are -1.0 and the other 6 elements are random values which are $-2.0 \sim +2.0$. ψ_{min} of each set is calculated by the way described in Sec.5.1.

Next, the convergence properties of these sets are investigated with various λ. Eq.(14) is the theoretical condition that the learning must converges regardless of the initial weights. Therefore, it is necessary to investigate the convergence properties using many different initial weights. For this reason, 100 random initial weights are used. As there are $5! = 120$ orders for pattern presentation, the learnings of 120 orders about each initial weights are investigated. That is, the convergence property about a certain pattern set and a certain λ is judged by 12000 trials.

Figure 9 shows the theoretical condition and the simulation results. Two solid lines represent the upper and the lower bounds of the theoretical condition for the convergence given by Eq.(14). Each vertical line of "o"and "x"means the convergence properties of a certain pattern set. "o"means that all the 12000 trials have converged. "x"means that the 12000 trials include which has not converged.

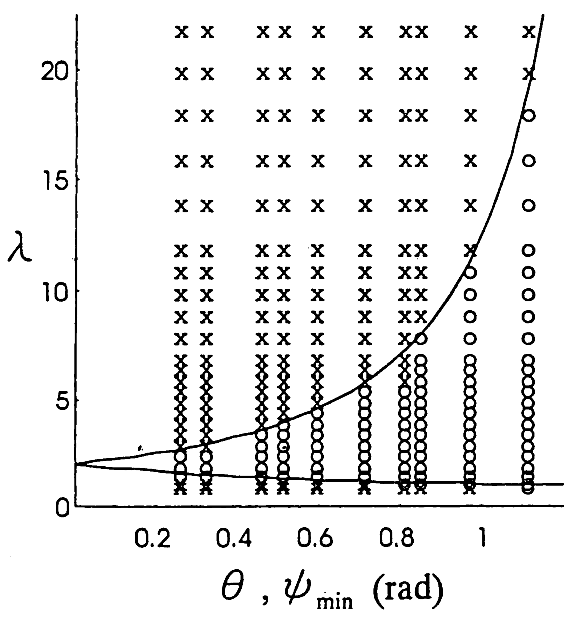

Fig. 9: Condition for 1-GLA convergence.

From Fig.9, the results of this simulation are summarized as follows:

- There is the slight difference between the theoretical condition and the simulation results. This fact indicates that the angle of the solution area ψ_{min} doesn't determine the convergence property exactly.
- The simulation results agree quite well with the theoretical condition. This fact indicates that the angle of the solution area ψ_{min} is a good estimation of the convergence property.

6. Conclusion

The geometric learning algorithm (GLA) has been proposed for an elementary perceptron. The condition for the convergence about the 1st order GLA for 2 patterns has been theoretically derived. The angle of the solution area ψ_{min} has been introduced and the meaning as the estimation of the convergence property has been investigated through computer simulation. It has been shown that ψ_{min} is a good estimation of the convergence property.

Acknowledgement

We thank Dr.Kazushi Ikeda for helpful discussions.

References

[1] S.Haykin. *Neural Networks —A Comprehensive Foundation—*, Macmillan College Publishing Company, 1994.

[2] K.Ozeki and T.Umeda. An adaptive filtering algorithm using an orthogonal projection to an Affine subspace and its properties (in Japanese), *IECE Trans.*, vol.J67–A, no.2, pp. 126–132, 1984.

[3] M.Hattori and M.Hagiwara. Intersection learning for bidirectional associative memory (in Japanese), *T.IEE Japan*, vol.116–C, no.7, pp. 755–761, 1996.

[4] F.Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review*, vol.65, pp. 386–408, 1958.

[5] B.Widrow and M.E.Hoff,Jr. Adaptive switching circuits, *IRE WESCON Conv. Rec.*, pt.4, pp. 96–104, 1960.

[6] J.I.Nagumo and A.Noda. A learning method for system identification, *IEEE Trans. on Automatic Control*, vol.AC–12, pp. 282–287, 1967.

[7] A.E.Albert and L.S.Gardner,Jr. Stochastic Approximation and Nonlinear Regression, *MIT Press*, 1967.

[8] A.Weiss and D.Mitra. Digital adaptive filters: conditions for convergence, rates of convergence, effects of noise and errors arising from the implementation, *IEEE Trans. Inf. Theory*, vol.IT–25, pp. 637–652, 1979.

[9] T.C.Hsia. Convergence analysis of LMS and NLMS adaptive algorithms, *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing*,Boston, Massachusetts, pp. 667–670, 1983.