1997 International Symposium on Nonlinear
Theory and its Applications (NOLTA'97)
Honolulu, U.S.A., Nov.29 – Dec. 2, 1997

Conditions for Convergence of the Normalized LMS Algorithm in Neural Learning

Kazushi IKEDA* Seiji MIYOSHI†‡ Kenji NAKAYAMA‡

* *Fac. Engineering, Kanazawa Univ.*
† *Kobe City College of Technology*
‡ *Sch. Natural Science and Technology, Kanazawa Univ.*

Abstract— The Perceptron learning algorithm for linear dichotomies can be regarded as the LMS algorithm which is one of the most popular algorithms for transversal filters. The normalized LMS (N-LMS) algorithm is one of the improved versions of the LMS algorithm for transversal filters and we apply it to linear dichotomies. In this paper, the proof of the convergence of the N-LMS algorithm for linear dichotomies in a finite number of iterations when the learning coefficient μ is unity, and the sufficient condition of μ for the convergence are given.

I. Introduction

In the field of adaptive filters, a linear filter called a transversal filter which outputs

$$y = x^t w \in R \tag{1}$$

where x and w are the input and weight vectors of the filter, respectively, is most widely used. The least mean square (LMS) algorithm is the most popular for a transversal filters and it changes the weight vector s much as

$$\Delta w = \mu x(d - x^t w), \tag{2}$$

where μ and d are the learning rate and the desired output, respectively, and then $(d - x^t w)$ means the output error. The LMS algorithm converge in probability when and only when $0 < \mu < \lambda_{\max}$ where $\lambda_{\max}$ is the largest eigenvalue of the auto-correlation matrix Σ[1], that means that we cannot say the learning rate which guarantees the convergence when Σ is unknown. The improved version at this point is the normalized LMS (N-LMS) algorithm shown as

$$\Delta w = \mu x(d - x^t w)/\|x\|^2, \tag{3}$$

where Δw does not depend on $\|x\|$ and the convergence condition is simply $0 < \mu < 2$[1, 3]

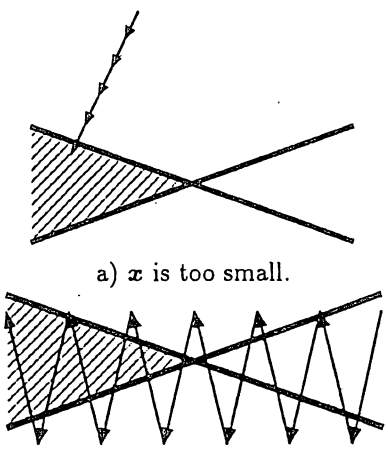
A linear dichotomy called a Perceptron

$$y = \text{sign}\,[x^t w], \tag{4}$$

on the other hand, consists of a transversal filter and a sign function sign $[\cdot]$ and is often used as an element of neural networks. Since the Perceptron learning (PL) algorithm is written as

$$\Delta w = \frac{1}{2} x \left(\text{sign}\,[x^t w_o] - \text{sign}\,[x^t w] \right) \tag{5}$$

where w_o is the true parameter, the PL algorithm can be regarded as an application of the LMS algorithm to a linear dichotomy. Because the PL algorithm ignores the magnitudes of x and w, its convergence speed is sometimes slow (Fig.1), though it is guaranteed in a finite number of iterations[4, 5]. In Fig.1, the shadowed



a) x is too small.

b) x is too large

Fig.1 Perceptron Learning Algorithm in Parameter Space

parts means the set of parameters which output the true sign for the given examples and the arrows show the changes of the weight vector. From the analogy of the LMS algorithm and the N-LMS algorithm, we can improve the PL as

$$\Delta w = -2\mu x x^t w/\|x\|^2 \tag{6}$$

which we call the N-LMS algorithm for linear dichotomies. As shown in Fig.2, the algorithm above becomes the projection onto and the symmetry with the

hyperplane $x^t w = 0$ when $\mu = 1/2$ and $\mu = 1$, respectively, that means that the N-LMS algorithm for linear

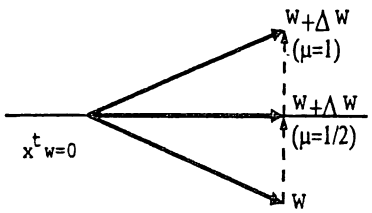


Fig.2 Geometrical view of the N-LMS algorithm

dichotomies is a special case of the orthogonal projection algorithm or the geometric learning algorithm[2] where the block signal processing techniques are applied to the orthogonal projection algorithm.

The Perceptron learning algorithm is guaranteed to stop in a finite number of iterations by Perceptron Convergence Theorem[4, 5]. On the other hand, the conditions for the convergence of the N-LMS algorithm has been little studied though the difference of their convergence conditions is worth being studied because it shows the effects of the normalization of the input vectors. It also elucidates the influences of the nonlinear element sign[·] by comparing with the condition of transversal filters' case.

Another property of the N-LMS algorithm for linear dichotomies is that it is free from the effects of the normalization of the weight vector. If the weight vector normalization is applied to the PL algorithm, its convergence is no longer guaranteed as shown in Fig.3. In this case, the weight vector A changes to B

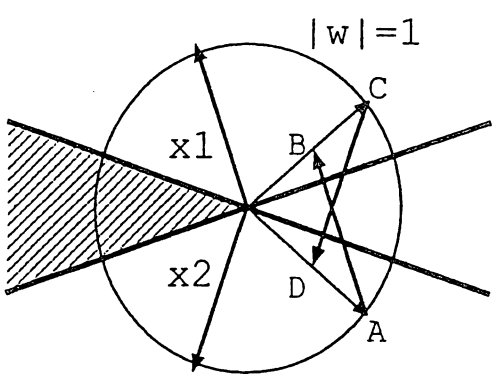


Fig.3 An example that the PL algorithm does not converge (in case of 2-D inputs and 2 given examples)

by one example x_1 and then to C by the normalization. Next, the other example x_2 carries C to D and the normalization makes D to return to A.

The problem about the convergence conditions of the N-LMS algorithm for linear dichotomies has been studied for the special case of 2-D inputs and 2 given examples[2], however, the results can not be straightforwardly applied to general cases. This paper proves that the N-LMS algorithm stops in a finite number of iterations when the learning rate $\mu = 1$ and gives a sufficient condition of μ for convergence.

II. The N-LMS Algorithm Convergence Theorem

Compared with that the optimal weight of a transversal filter is a point where the mean squared error is minimized, the set any weights in which output the same signs as the true weight has area which we call admissible the consistent area. This is the largest difference between them and it affects their convergence properties of the N-LMS algorithms. The weight of a transversal filter gradually approaches to the optimal point when $0 < \mu < 2$ but never reaches to the optimal because its convergence is probabilistic. The weight of a linear dichotomy, on the other hand, reaches the consistent area and stops at there when $\mu = 1$. We prove it in the following.

We denote the ith given example *i.e.* the pair of the input vector and the output by x_i and y_i, respectively. Since the true output for $-x_i$ is $+1$ when that for x_i is -1, we assume that the output y_i is always positive and call x_i itself the ith example. And the magnitude of x_i does not affect the learning, $\|x_i\| = 1$ is also assumed. Then, the domain of x_i becomes a half of $m-1$ dimensional hypersphere S^{m-1}

$$S_+ = \left\{ x_i | x_i^t w_o > 0, x \in S^{m-1} \right\}. \qquad (7)$$

where w_o is the true weight vector.

When p examples $x_i, i = 1, \ldots, p$ are given, the consistent area D is defined as

$$D = \left\{ w | x_i^t w > 0, i = 1, \ldots, p \right\}. \qquad (8)$$

Any weight w in D gives the same output as w_o for x_i. Because an example x_i which satisfies $x_i^t w < 0$ is used for the learning, the learning stops when $w \in D$. Now we start to prove that the N-LMS algorithm stops in a finite number of iterations when $\mu = 1$ using the mathematical induction. Since w moves symmetrically with the hyperplane $x_i^t w = 0$ when $\mu = 1$ and x_i is given, the magnitude of w does not change. So, we assume that $\|w\|$ is always unity, therefore, $w \in S^{m-1}$. Then, D is a polyhedron on S^{m-1}. We denote a weight in D by w^* and the angle between w^* and x_i by $\pi/2 - \theta_i$, respectively.

First, we prove that in the 2-dimensional case of w. As shown in Fig.4, the learning by x_i makes the angle between $w + \Delta w$ and w^* $2\theta_i$ smaller than that between w and w^*. Therefore, the learning stops in a finite number of iterations.

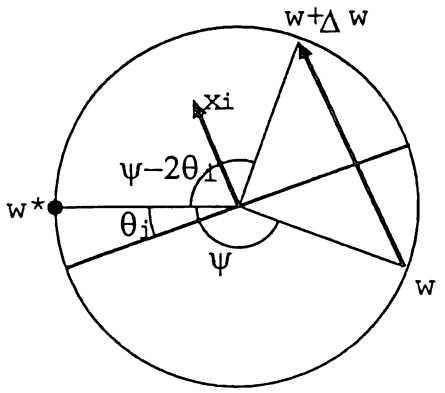


Fig.4 The N-LMS algorithm (2-dimensional case)

Next, we assume that the theorem is true when $m \leq k$ and show that it holds even when $m = k+1$ using reductive absurdity. It is assumed that $\boldsymbol{w}$ does not enter D in a finite number of iterations. Then, there is at least one example which must be applied infinite times, which we denote by $\boldsymbol{x}_i, i = 1, \ldots, p'$. Because $\boldsymbol{x}_i, i = p'+1, \ldots, p$ are applied only a finite times, we consider that the learning by them already finished and only $\boldsymbol{x}_i, i = 1, \ldots, p'$ are used. The learning by $\boldsymbol{x}_i$ whose angle with $\boldsymbol{w}^*$ is $\pi/2 - \theta_i$ increases $\boldsymbol{w}$'s component of $\boldsymbol{w}^*$ direction $\varepsilon \sin\theta_i$. Since $\boldsymbol{w}$ is constant, this means that the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$ decreases. So, ε has to approach to 0 in order that $\boldsymbol{w}$ does not enter the consistent area D, *i.e.* $\boldsymbol{w}(\infty)$ must be on $\boldsymbol{x}_i^t \boldsymbol{w} = 0$. Therefore, $X = \mathrm{span}(\boldsymbol{x}_i, i = 1, \ldots, p') = R^m$, $m < k+1$ because $\boldsymbol{w}(\infty) = 0_{k+1}$ contradicts $\|\boldsymbol{w}\| = 1$ if $X = R^{k+1}$. Since the N-LMS algorithm adds $\boldsymbol{w}$ a vector proportional to $\boldsymbol{x}_i$, it does not change $\boldsymbol{w}$'s component orthogonal to X. So, we divide $\boldsymbol{w}$ to $\boldsymbol{w}' + \boldsymbol{w}^\perp$ where $\boldsymbol{w}' \in X$ and $\boldsymbol{w}^\perp \perp X$. We denote the consistent area in X which made by $\boldsymbol{x}_i, i = 1, \ldots, p'$ by D'. So, since X is a k or less dimensional space, $\boldsymbol{w}'$ enter D' in a finite number of iterations and then $\boldsymbol{x}_i^t \boldsymbol{w}' > 0$ for any i. Since $\boldsymbol{x}_i^t \boldsymbol{w} = \boldsymbol{x}_i^t \boldsymbol{w}', i = 1, \ldots, p'$ and then $\boldsymbol{x}_i^t \boldsymbol{w} > 0, i = 1, \ldots, p'$, $\boldsymbol{x}_i, i = 1, \ldots, p'$ can not be used for learning any more. From the assumption that the learning by $\boldsymbol{x}_i, i = p'+1, \ldots, p$ finishes, the learning has to stop, which contradicts the assumption that $\boldsymbol{w}$ does not enter D in a finite number of iterations.

III. Conditions of the Learning Rate for Convergence

In the previous section, the N-LMS algorithm converges when $\mu = 1$. Then, does μ have to be unity? In this section, we give a sufficient condition of μ for convergence. It is shown as an interval which includes unity but depends on the given examples, which means that $\mu = 1$ is best for applications.

If the learning by $\boldsymbol{x}_i$ decreases the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^* \in D$, the convergence in a finite number of iterations can be shown in the same way as $\mu = 1$. So, we derive the interval of μ where the angle necessarily decreases. We separately consider the 2-dimensional space W spanned by $\boldsymbol{w}^*$, $\boldsymbol{x}_i$ and its complement $W^\perp$. When $\boldsymbol{w} \in W$, consider a circle whose radius is $\|\boldsymbol{w}\|$

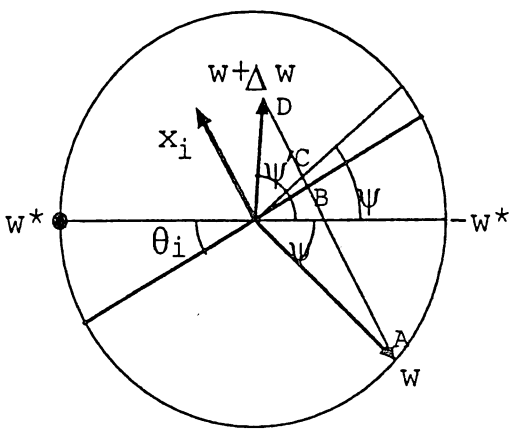


Fig.5 View in the case that $\boldsymbol{w} \in \mathrm{span}(\boldsymbol{w}^*, \boldsymbol{x}_i)$

and define ψ and ψ' as the angle between $-\boldsymbol{w}^*$ and $\boldsymbol{w}$ and that between $-\boldsymbol{w}^*$ and $\boldsymbol{w} + \Delta\boldsymbol{w}$, respectively, as shown in Fig.5. When $-\theta_i < \psi \leq 0$, if $\mu > 0$ then $\psi' > 0$ and $\psi' > \psi$. When $0 < \psi < \pi/2 - \theta_i$, if the terminus D of the vector $\boldsymbol{w} + \Delta\boldsymbol{w}$ is on the half line AC then $\psi' > \psi$ holds. Here, consider that AD is proportional to μ and D coincides with B when $\mu = 1/2$, and then it is derived that $\psi' > \psi$ if $\mu/(1/2)$ is bigger than AC/AB, that is,

$$\left[\sin(\psi+\theta_i) + \cos(\psi+\theta_i)\tan(\psi-\theta_i)\right] / \sin(\psi+\theta_i) < \frac{\mu}{\frac{1}{2}}. \tag{9}$$

So, the condition of μ for $\psi < \psi'$ for any ψ is Therefore, if $\theta_i < \pi/4$ then

$$\mu > \frac{1}{2}\left[1 + \frac{\tan(\pi/4 - \theta_i)}{\tan(\pi/4 + \theta_i)}\right] \tag{10}$$

$$= \frac{1 + \tan^2\theta_i}{(1 + \tan\theta_i)^2} \tag{11}$$

and if $\pi/4 \leq \theta_i < \pi/2$ then $\mu > \frac{1}{2}$ In the same way, it is derived that if $\theta_i < \pi/4$ then

$$\mu < \frac{1 + \tan^2\theta_i}{(1 - \tan\theta_i)^2} \tag{12}$$

and if $\pi/4 \leq \theta_i < \pi/2$ then $\mu > 0$. Because the conditions above has to hold for any $\boldsymbol{x}_i$, they can be concluded that if $\theta^* < \pi/4$ then

$$\frac{1 + \tan^2\theta^*}{(1 + \tan\theta^*)^2} < \mu < \frac{1 + \tan^2\theta^*}{(1 - \tan\theta^*)^2}. \tag{13}$$

and if $\pi/4 \leq \theta^* < \pi/2$ then $\mu > \frac{1}{2}$ where w^* is such that maximizes $\min_i \theta_i$ and θ^* is the minimum. We call the above Condition I

Next, we consider the case $w \notin W$. Instead of the angle between a vector and w^*, We evaluate a normalized vector's component of w^* direction, *i.e.* the increase of the component is equivalent to the decrease of the angle. w can be divided as $w = w' + w^\perp$ where $w' \in W$ and $w^\perp \perp W$ and then the learning by x_i does not change $w^\perp$. Therefore, the projections of w and $w + \Delta w$ onto W are shown as in Fig.6 where the center and the radius of the circle are the origin and $\|w'\|$, respectively. When $\mu < 1$, since $w' + \Delta w$ denoted by B

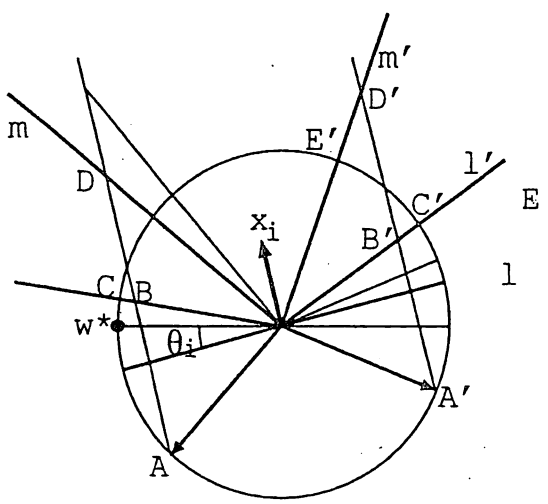


Fig.6 View in the case that $w \notin$ span (w^*, x_i)

(or B') exists within the circle and $\|w'\| > \|w' + \Delta w\|$, the component of $W^\perp$ in $\frac{\|w\|}{\|w+\Delta w\|}(w + \Delta w)$ whose magnitude is normalized as much as that of w is more than that of $w + \Delta w$. Therefore, its projection onto W is on l (or l'), more precisely, between B (or B') and C (or C'). If μ satisfies Condition I, the angle between l (or l') and w^* necessarily decreases, the component of w^* direction in any point on BC (or $B'C'$) increases, that is, the angle between $w + \Delta w$ and w^* is less than that between w and w^*. In the same way, the above is easily proven even in the case of $\mu > 1$.

Because Condition I depends on an angle θ^* which is given by the examples and a parameter w^* in the consistent area, the learning rate μ must be 1 from the practical point of view.

IV. Conclusion

From the analogy of the Perceptron learning algorithm and the LMS algorithm, the N-LMS algorithm can be applied to linear dichotomies. This paper gives its convergence properties, especially proves that the algorithm stops in a finite number of iterations when $\mu = 1$. Besides, a sufficient condition for convergence is given which is an interval which includes unity. The results are contrastive with the transversal filter's case ($0 < \mu < 2$) or the perceptron's case (any positive number).

References

[1] Haykin, S.: *Adaptive Filter Theory*, Prentice-Hall, 2nd edition, 1991.

[2] Miyoshi, S. and Nakayama, K.: Geometric Learning Algorithm for Elementary Perceptron, *Proc. Int'l Conf. Neural Networks*, 1997, 1913–1918.

[3] Nagumo, J. and Noda, A.: A Learning Method for System Identification, *IEEE Trans. AC*, Vol. 12 (1967), 282–287.

[4] Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychological Review*, Vol. 65 (1958), 386–408.

[5] Rosenblatt, F.: *Principle of Neurodynamics*, Spartan, Washington, D. C., 1961.