ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 25-30

Stability Bounds of Momentum Coefficient and Learning Rate in Backpropagation Algorithm

Ziqiang Mao and T. C. Hsia

Department of Electrical and Computer Engineering
University of California at Davis, CA, 95616

Abstract: In this paper we rigorously derive and prove the stability bounds of the momentum coefficient α and the learning rate η of the backpropagation updating rule in multilayer neural networks. The bounds of α is found to be $-1<\alpha<+1$ rather than $0<\alpha<1$ as stated in most literature. The theoretical upper bound of η is derived and its practical approximation is obtained, which can serve as a convenient guide for choosing η in applications. It is shown that the upper bound of η is proportional to $1+\alpha$. These properties are verified in simulation studies of the XOR problem.

1.Introduction

Backpropagation algorithm is a popular updating rule for multilayer neural networks and its performance has been widely reported in the literature for various applications. The backpropagation algorithm was derived using the gradient descent method. To achieve stability, it is well known in practice that the learning rate η in the algorithm should be positive and the momentum coefficient α should be in the interval [0,1). However the stability bounds of α and η have never been rigorously derived. Thus it is of great interest to neural network researchers to know the stability bounds of these two parameters when applying the backpropagation algorithm. Otherwise, one has to resort to a trial-and-error method to find a pair of α and η which are stable. In this paper we present a stability analysis of backpropagation algorithm which establishes the necessary conditions of stability for α and η. The main results are (1)The stability bounds of α is $-1<\alpha<1$ rather than $0\le\alpha<1$; (2)The bounds for η is $0<\eta<B(k,\alpha)$, where $B(k,\alpha)$ is a time-varying upper bound dependent on α; (3)An computable approximation of $B(k,\alpha)$ is given as $2(1+\alpha)b(k)$, where $b(k)$ can be computed at each iteration step k; (4)These results are verified in simulation of the XOR problem.

This paper is organized as follows. In Section 2 we describe the neural network and the weight updating rule. The stability bounds of α are proven in section 3. The bounds of η are derived in Section 4. In Section 5, we present an approximation to the theoretical upper bound of η which is easily computable. Simulation study on the XOR problem is presented in Section 6. Section 7 provides the conclusion.

2.Neural Network and Updating Rule

Let us define the following notations:

p: pattern index, p=1,2,...,L;
L: number of patterns;
m: dimension of input of a neural network;
n: dimension of output of a neural network;
N: number of all weights;

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 25-30

k: weight updating step;
w: vector containing all weights in a multilayer neural network;
x_p: input pattern of a neural network;
y_p: output of a neural network;
$f(.)$: function formed by a multilayer neural network, $y_p=f(w, x_p)$;
$f_i(.)$: element of vector $f(.)$;
d_p: desired output of a neural network for the input x_p;
w^*: an equilibrium point of the weight vector such that $d_p=f(w^*, x_p)$.

A multilayer neural network can be used to approximate an unknown nonlinear mapping whose inputs and outputs are observable. The original theoretical basis for the approximation is the Kolmogorov[1957] theorem. It has been proven by Cybenko[1989] that any continuous function can be approximated by a two-layer neural network to any desired accuracy. The weight adaptation method is the back-propagation algorithm independently proposed by Werbos[1974], Parker[1985] and Rumelhart[1986]. For an objective function $E(w)$ to be minimized, the updating rule for adjusting weights in neural network is expressed by

$$w(k+1) = w(k) + \eta(-\frac{dE}{dw}) + \alpha(w(k)-w(k-1)) \quad (1)$$

where η is the learning rate and α is the momentum coefficient. It is well known in the literature (Rumelhart[1986]) that η is positive and α is in $[0,1)$. No proof is given however.

Suppose that the structure of a multilayer feedforward neural network is determined and it can represent the unknown nonlinear mapping exactly with proper weights w^*. That is, for a given input pattern x_p, the desired output d_p can be obtained from the neural network output $y_p=f(w,x_p)$ with weight $w=w^*$, or $d_p=f(w^*, x_p)$.

The objective function $E(w)$ is usually defined as the sum of squared error between the actual output y_p of neural network and the desired output d_p for all training patterns:

$$E = \frac{1}{2}\sum_{p=1}^{L}(d_p-y_p)^T(d_p-y_p) \quad (2)$$

Then the updating rule (1) will have the following form:

$$w(k+1) = w(k) + \eta \sum_{p=1}^{L} M_p(k)^T e_p(k) + \alpha(w(k)-w(k-1)) \quad (3)$$

where $M_p(k) = \frac{df(w(k),x_p)}{dw(k)}$ and $e_p(k) = d_p-y_p(k)=f(w^*, x_p) - f(w(k), x_p)$.

In this paper, we will derive the necessary conditions of η and α which ensure the stability and convergence of the updating rule (3).

3.Bounds of Momentum Coefficient α

Updating rule (3) can be written as

$$w(k+1) = w(k) +\eta s(k)+\alpha(w(k)-w(k-1)) \quad (4)$$

where $s(k) = \sum_{p=1}^{L} M_p(k)^T e_p(k)$.

Defining $v(k)=w(k)-w(k-1)$, we get from (4)

$$v(k+1) = \alpha v(k) +\eta s(k) \quad (5)$$

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 25-30

The solution of (5) is

$$v(k+1) = \alpha^k v(1) + \eta \sum_{i=1}^{k} \alpha^{k-i} s(i) \qquad (6)$$

Recalling the assumption that $d_p = f(w^*, x_p)$, we get following Lemma:

Lemma 1. If $w(k) \to w^*$, then $v(k) \to 0$ and $e_p(k) \to 0$.

Proof: $v(k) = w(k) - w(k-1) \to 0$ as $w(k) \to w^*$. $e_p(k) = d_p - y_p(k) = f(w^*, x_p) - f(w(k), x_p)$, and $f(w,x)$ is a continuous function of w. Hence, $e_p(k) \to 0$ for all $p=1,2,...,L$, as $w(k) \to w^*$. QED

Theorem 1. A necessary condition for the convergence of the updating rule (3) is that $|\alpha|<1$.

Proof: By Lemma 1, $v(k) \to 0$ and $e_p(k) \to 0$ as $w(k) \to w^*$. That is, for any $\varepsilon>0$, there exists an integer K such that $\|v(k)\|<\varepsilon$ and $\|e_p(k)\|<\varepsilon$ for all $k>K$. For $k=K+1$, Eq.(6) can be rewritten as:

$$\alpha^{K+1} v(1) + \eta \sum_{i=1}^{K} \alpha^{K+1-i} s(i) = v(K+2) - \eta s(K+1) \qquad (7)$$

In view of (7), the following inequality can be easily derived:

$$|\alpha|^{K+1} \|v(1) + \eta \sum_{i=1}^{K} \alpha^{-i} s(i)\| \le \|v(K+2)\| + \eta \|s(K+1)\| \qquad (8)$$

As $w(K+1) \to w^*$, $v(K+2) \to 0$, and $s(K+1) \to 0$ because $e_p(K+1) \to 0$, $w(K+1)$ is bounded, and all elements of $M_p(K+1)$ are bounded. Therefore, the right hand side of (8) goes to zero. So the left hand side also goes to zero.

Howeve, $\|v(1) + \eta \sum_{i=1}^{K} \alpha^{-i} s(i)\|$ cannot be zero in general, therefore, $|\alpha|^{K+1}$ should go to zero as $K \to 0$. Consequently $|\alpha|<1$ is required. QED

This theorem shows that $|\alpha| <1$ is a necessary condition. If $|\alpha| \ge 1$, then the updating rule is unstable. The momentum coefficient α is thus bounded by $-1<\alpha<1$.

4.Theoretical Lower and Upper Bound of Learning Rate η

We will establish the bounds for η in two cases: $\alpha=0$ and $0<|\alpha|<1$.

4.1 The case $\alpha = 0$

When $\alpha=0$, (3) can be written as

$$w(k+1) = w(k) + \eta s(k) \qquad (9)$$

Define $u(k)=w^*-w(k)$, (9) is the same as $u(k+1) = u(k) - \eta s(k)$. One way to show the convergence of $w(k)$ is to require that $\|u(k+1)\|^2 \le \|u(k)\|^2$. Then the following inequality must be true:

$$\|u(k+1)\|^2 = \|u(k)\|^2 + \eta^2 s^T(k)s(k) - 2\eta u^T(k)s^T(k) \le \|u(k)\|^2 \qquad (10)$$

That is

$$\eta^2 s^T(k)s(k) \le 2\eta u^T(k)s(k) \quad \text{for all } k. \qquad (11)$$

Lemma 2. Let $B(k,\alpha) = \dfrac{2u^T(k)s(k)}{s^T(k)s(k)}$ when $\alpha=0$, then $\lim_{w(k) \to w^*} B(k,0) = 2b$, where

$$b=\frac{\sum_{i,j=1}^{n} q_{ij}}{n\sum_{i,j=1}^{n} r_{ij}},\quad [q_{ij}]_{nxn}=\overline{M}M^T,\quad [r_{ij}]_{nxn}=MM^T,\quad M=\sum_{p=1}^{L} M_p(w^*),$$

$$M_p(w^*)=\frac{df(w^*,x_p)}{dw^*},\quad \overline{M}=\sum_{p=1}^{L}\overline{M}_p(w^*),\ \text{and}\ \overline{M}_p(w^*)=\left[\left(\frac{\partial f_i(w^*,x_p)}{\partial w_j}\right)^{-1}\right]_{nxn}.$$

Proof: For simplicity, we only prove the case with L=1.

$$\lim_{w(k)\to w^*}\frac{u^T(k)M^T(k)e(k)}{e^T(k)M(k)M^T(k)e(k)}$$

$$=\lim_{w(k)\to w^*}\frac{e^T(k)h(k)u^T(k)M^T(k)e(k)}{n\ e^T(k)M(k)M^T(k)e(k)}=\frac{\sum_{i,j=1}^{n} q_{ij}}{n\sum_{i,j=1}^{n} r_{ij}}=b, \qquad (12)$$

where $h(k)=[(f_1(w^*, x)-f_1(w(k), x))^{-1}, ...,(f_n(w^*, x)- f_n(w(k), x))^{-1}]^T$. QED

Theorem 2. A necessary condition for the convergence of the updating rule (9) is that η be positive.

Proof: When $\eta=0$, the updating rule is trivial because there will not be any weight updating in (9). When η is negative, we will show it is impossible by contradiction. For the simple case of n=1 and L=1, the inequality (11) becomes (13) when both sides of (11) are divided by η,

$$\eta e^T(k)M(k)M^T(k)e(k)\geq 2u^T(k)M^T(k)e(k) \qquad (13)$$

where $M(k)=\frac{df(w(k),x)}{dw(k)}$ is a 1 x N vector, and $e(k) = d-y(k)=f(w^*, x) - f(w(k), x)$ is a scalar. Then from (13), we get

$$\eta \geq \frac{2u^T(k)M^T(k)e(k)}{e^T(k)M(k)M^T(k)e(k)} \qquad (14)$$

By Lemma 2, after taken the limit $w(k)\to w^*$ on both sides of (9), (14) becomes

$$\lim_{w(k)\to w^*}\eta \geq \frac{2N}{MM^T} \qquad (15)$$

where $M=\frac{df(w^*,x_p)}{dw^*}$. That means $\eta\geq 0$, which contradicts the assumption that η is negative. Therefore, η must be positive. QED

Theorem 2 claims that the lower bound of η is zero when $\alpha=0$. Given that $\eta>0$, the upper bound of η for $\alpha=0$ can be derived from (11) as follows:

$$\eta \leq \frac{2u^T(k)s(k)}{s^T(k)s(k)} = B(k,\alpha) \qquad \text{when } \alpha=0. \qquad (16)$$

4.2.The case $0<|\alpha|<1$

Making use of the definition $u(k) = w^* - w(k)$, we get from (3)

$$u(k+1) = u(k) -\eta s(k)+\alpha(u(k) - u(k-1)) \qquad (17)$$

Multiplying both sides of (17) by $s^T(k)$ yields

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 25-30

$$\eta s^T(k)s(k) = (1+\alpha)s^T(k)u(k) - \alpha s^T(k)u(k-1) - s^T(k)u(k+1) \quad (18)$$

Furthermore we can obtain the following inequality from (18)

$$|\eta| s^T(k)s(k) \le (1+\alpha)|s^T(k)u(k)| + |\alpha s^T(k)u(k-1) + s^T(k)u(k+1)| \quad (19)$$

Therefore, the upper bound of $|\eta|$ is given by

$$|\eta| \le B(k,\alpha) \quad \text{for } 0<|\alpha|<1 \quad (20)$$

where $B(k,\alpha)=\dfrac{(1+\alpha)|s^T(k)u(k)|+|\alpha s^T(k)u(k-1)+s^T(k)u(k+1)|}{s^T(k)s(k)}$, when $0<|\alpha|<1$.

When $\eta=0$, (3) becomes $w(k+1)=(1+\alpha)w(k)-\alpha w(k-1)$, which is a system with eigenvalues 1 and α. The system is marginally stable. Hence, η should be positive for $0<|\alpha|<1$.

The above results can be summarized by the following theorem:

Theorem 3. A necessary condition for the convergence of updating rule (3) is that η should be $0<\eta<B(k,\alpha)$ for $-1<\alpha<1$. **QED**

5.Computable Upper Bound of Learning Rate η

Theorem 3 gives the theoretical upper bounds of η which cannot be easily evaluated because w^* is not known a priori so that $u(k) = w^* - w(k)$ cannot be computed as required by $B(k,\alpha)$. Thus it would be helpful for us to find computable bounds which are approximations to the theoretical bounds.

By Lemma 2, as $w(k)\text{-->}w^*$, the limit of $B(k,\alpha)$ is $2b$ for $\alpha=0$, and the limit of $B(k,\alpha)$ is $2b(1+\alpha)$ for $0<|\alpha|<1$. Because $2b(1+\alpha)$ is equal to $2b$ when $\alpha=0$, we can say that for all α in $(-1,+1)$, the limit of $B(k,\alpha)$ is $2b(1+\alpha)$ at convergence.

Note that

$$\lim_{w(k)\to w^*} b(k) = b \quad (21)$$

where $b(k)=\dfrac{\sum_{i,j=1}^{n} q_{ij}(k)}{\sum_{i,j=1}^{n} r_{ij}(k)}$, $[q_{ij}(k)]_{nxn}=\overline{M}(k)M^T(k)$, $[r_{ij}(k)]_{nxn}= M(k)M^T(k)$,

$M(k)=\sum_{p=1}^{L} M_p(k)$ and $\overline{M}(k)=\sum_{p=1}^{L} \overline{M}_p(w(k))$.

Hence, we can use $b(k)$ as an approximation to b at each iteration step k in the learning process. In summary, we have the following theorem:

Theorem 4. For all α in $(-1,+1)$, an approximation of the upper bound of η is $\min_{k\ge 1}\{2b(k)(1+\alpha)\}$. The upper bound can be computed at each step k as the weights are updated. **QED**

Theorem 4 states that the upper bound of η is proportional to $1+\alpha$. Since the upper bound is only a necessary condition, a more conservative value of η should be used in applications.

6.Simulation Results

The exclusive OR (XOR) problem is investigated here to verify the results obtained above. A two-layer neural network with 2 neurons at hidden layer can represent the XOR function as shown by Rumelhart[1986]. With the biases added, there are 9 weights to be adjusted. The activation function at the hidden layer is a sigmoid function $y=1/(1+e^{-x})$. The activation function at the output layer is linear $y = x$.

The update rule Eq.(3) is applied for a range of α values, and for each α, a small η is chosen to insure convergence. As shown in Table 1, 9 simulation runs were made. For each case, the number of iterations (NI) required to reduce the error E in Eq.(2) below 0.001 is given. In addition the upper bound $2b(k)(1+\alpha)$ for each case was computed at each iteration, and the minimum value,denoted as Bc, is listed. It is noted that the minimum values for all cases occur at convergence. The data of α and Bc in Table 1 satisfy the relation $Bc= 0.445+0.434\alpha \approx 0.44(1+\alpha)$ as predicted by Theorem 4. This bound Bc allows us to choose any pair (α,η) which guarantees convergence.

Table 1.Computed Bounds Bc vs α

α	-0.9	-0.7	-0.5	-0.2	0.0	0.2	0.5	0.7	0.9
η	0.02	0.05	0.08	0.13	0.16	0.2	0.25	0.28	0.31
NI	24843	8892	4905	2416	1638	1050	531	293	124
Bc	0.045	0.14	0.23	0.36	0.45	0.54	0.67	0.76	0.81

On the other hand, we can experimentally determine the actual stability bounds of η for different α in the following way. For each α in the range (-0.9,0.9), we ran the neural network program with different learning rates η. The maximum value of η that can be used before the objective function E starting to diverge is taken as the actual stability upper bound denoted as Ba. The results so obtained are shown in Table 2. These Ba data satisfy the relation $Ba= 0.49+0.5\alpha \approx 0.49(1+\alpha)$. Comparing Bc and Ba, we see that Bc is close to Ba, and yet Bc is more conservative. Thus Bc is a practical and easily computable upper bound for η.

Table 2.Actual Bounds Ba vs α

α	-0.9	-0.7	-0.5	-0.2	0.0	0.2	0.5	0.7	0.9
Ba	0.04	0.14	0.24	0.39	0.48	0.59	0.74	0.84	0.94

7.Conclusions

Stability bounds for the momentum coefficient α and the learning rate η for the backpropagation algorithm have been derived. These bounds are $-1<\alpha<1$ and $0<\eta<\min_{k\geq 1}\{2b(k)(1+\alpha)\}$. The validity of these bounds are verified in simulation studies of XOR problem.

References:

Cybenko,G.[1989]. "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol 2(4),pp. 303-314

Kolmogorov,A.N.[1957]. "On the representation of continuous function of many variables by superposition of continuous functions of one variable and addition," *Dolk. Akad. Nank*, USSR, vol.114, pp.953-956.

Parker, D.B.[1985]. "Learning Logic," *Tech. Rept. TR-47*, Center for Comput. Res. Econ. and Manage., MIT, Cambridge, MA.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J.[1986]. "Learning internal representations by error propagation," *Parallel Distributed Processing:Explorations in the Microstructure of Cognition*, vol.I, ed. D.E.Rumelhart *et al*, pp.318-363.

Werbos,P.[1974]. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," *Ph.D. Thesis*, Harvard Univ. Cambridge, MA.