Convergence of Online Gradient Method with Momentum for BP Neural Network

Chengyan Pan*, Gang Xie
School of Science, East China University of Science and Technology, Shanghai, China

*Corresponding author: Y30160132@mail.ecust.edu.cn

Abstract. The BP neural network, which uses the steepest descent method (gradient descent method) as the basic idea for learning, is often used to deal with approximation problems because of its strong nonlinear mapping ability. The gradient method with added momentum can improve the learning speed of BP neural network. We study the convergence of the online gradient method with momentum for two-layer BP neural network when the training samples are randomly arranged in each iteration. Choosing appropriate learning rates, and selecting momentum coefficients in an adaptive manner, we prove the weak and strong convergence theorems of the algorithm.

1. Introduction
Artificial neural network simulates biological neural network, has a unique model structure and strong nonlinear mapping ability, and is one of the major achievements in the history of artificial intelligence. The study of artificial neural network learning methods is also one of the research hotspots in recent years. In 1986, a team of scientists led by Rumelhart and McCelland proposed the BP neural network, which was later widely used [1-5]. As a multi-layer feedforward neural network learning based on the steepest descent method (gradient descent method), BP neural network is often used to deal with approximation problems. Some documents briefly discuss the convergence of two-layer BP neural network [6-8].

In order to accelerate and stabilize the learning rate, researchers have made different improvements to the online gradient method of the BP neural network. Among them, increasing the momentum term [9-12] and the penalty term [13-15] are two common methods. In his article [16], N.M. Zhang gave the convergence results of the online gradient method with momentum under certain conditions, it is necessary to assume that the error function is uniformly convex and requires that the training samples are linearly independent. Furthermore, he found that for the strong and weak convergence of the two-layer BP neural network, the training samples do not need to be linearly independent, and the error function does not need to be uniformly convex [17]. H. Shao et al. first added a penalty term to the online gradient method with momentum and gave the algorithm's strong and weak convergence results for the two-layer BP neural network [18], and further improved the penalty term and gave the results of the strong and weak convergence of the algorithm on the two-layer and three-layer BP neural network are shown [19].

N.M. Zhang’s article gives the convergence of the online gradient method, but in his algorithm, the sample order is always the same in the training process. In the actual training, we can consider randomly
arranging the training samples in each iteration cycle. This article mainly studies the above two training methods of the two-layer BP neural network. The online gradient method with momentum is introduced in Section 2, the main results are given in Section 3, the detailed proof is given in Section 4, and the summary is given in Section 5.

2. Online gradient method with momentum for two-layer BP neural network

Consider a two-layer BP neural network, as shown in Figure 1. If there are multiple neurons in the second layer, the weight and output of each neuron are independent of each other, so their training is also independent of each other. We can only consider the case where there is only one neuron in the second layer, and the conclusions we get also apply to networks with multiple neurons in the second layer.

Denote the weight vector by \( \mathbf{w} = (w_1, w_2, \cdots, w_n) \in \mathbb{R}^n \), and let \( g : \mathbb{R} \rightarrow \mathbb{R} \) be the smooth activation function of the neural network. For a given training set \( \{ (\mathbf{x}^j, o^j) \}_{j=1}^J \subseteq \mathbb{R}^n \times \mathbb{R} \), the actual output of the neural network is

\[
\mathbf{z}^j = g(\mathbf{w} \cdot \mathbf{x}^j), \ j = 1, \cdots, J
\]  

(1)

Where \( \mathbf{w} \cdot \mathbf{x}^j \) represents the inner product. The purpose of training the neural network is to select the weights such that the difference between the actual output and the target output is as small as possible, that is, the smaller \( |o^j - \mathbf{z}^j| \) is, the better. Generally, the objective function is selected as

\[
E(\mathbf{w}) = \frac{1}{2} \sum_{j=1}^J (o^j - \mathbf{z}^j)^2 = \frac{1}{2} \sum_{j=1}^J (o^j - g(\mathbf{w} \cdot \mathbf{x}^j))^2 .
\]  

(2)

In other words, the purpose of training is to find \( \mathbf{w}^* \in \mathbb{R}^n \), such that

\[
E(\mathbf{w}^*) = \min_{\mathbf{w} \in \mathbb{R}^n} E(\mathbf{w}) .
\]  

(3)

We solve the minimization problem with the gradient method, and the gradient of \( E \) with respect to \( \mathbf{w} \) is

\[
E_{\mathbf{w}}(\mathbf{w}) = -\sum_{j=1}^J (o^j - g(\mathbf{w} \cdot \mathbf{x}^j)) \cdot g'(\mathbf{w} \cdot \mathbf{x}^j) \cdot \mathbf{x}^j .
\]  

(4)
Define \( g_j(t) = \frac{1}{2} \left( \alpha^j - g(t) \right)^2 \), \( j = 1, \ldots, J \), then

\[
E(w) = \sum_{j=1}^{J} g_j(w \cdot \xi^j),
\]

\[
E_w(w) = \sum_{j=1}^{J} g'_j(w \cdot \xi^j) \cdot \xi^j.
\]

Now consider the online gradient algorithm that drives the amount. Here we disrupt the order of the training samples during training, that is, rearrange the training samples in each iteration cycle. Let \( \{\xi^1, \xi^2, \ldots, \xi^J\} \) be the random arrangement of \( \{\xi^1, \xi^2, \ldots, \xi^J\} \) in the mth cycle in iterative training. Given initial weights \( w^{-1} \) and \( w^0 \), update \( w \) through the following algorithm:

\[
w^{m+j} = w^{m+j-1} - \eta_m g_{mj} (w^{m+j-1} \cdot \xi^mj) \cdot \xi^mj + \tau_{mj} (w^{m+j-1} - w^{m+j-2}); \quad j = 1, \ldots, J; \quad m = 0, 1, \ldots
\]

Let

\[
\Delta w^{m+j-1} = w^{m+j-1} - w^{m+j-2}; \quad j = 1, \ldots, J; \quad m = 0, 1, \ldots
\]

Then

\[
w^{m+j} = w^{m+j-1} - \eta_m g_{mj} (w^{m+j-1} \cdot \xi^mj) \cdot \xi^mj + \tau_{mj} \Delta w^{m+j-1}; \quad j = 1, \ldots, J; \quad m = 0, 1, \ldots
\]

That is

\[
\Delta w^{m+j} = -\eta_m g_{mj} (w^{m+j-1} \cdot \xi^mj) \cdot \xi^mj + \tau_{mj} \Delta w^{m+j-1}; \quad j = 1, \ldots, J; \quad m = 0, 1, \ldots
\]

Where \( \eta_m \) represents the learning rate in the mth iteration cycle and \( \tau_{mj} \Delta w^{m+j-1} \) is so-called momentum term with the momentum coefficient \( \tau_{mj} \geq 0 \).

For the convenience of calculation, set

\[
p^{m,j} = g_{mj} (w^{m+j-1} \cdot \xi^mj) \cdot \xi^mj, \quad i, j = 1, \ldots, J; \quad m = 0, 1, \ldots
\]

Particularly, when \( i = 1 \)

\[
p^{m,j} = p^{m,j-1} = g_{mj} (w^{m+j} \cdot \xi^mj) \cdot \xi^mj, \quad j = 1, \ldots, J; \quad m = 0, 1, \ldots
\]

Then

\[
E_w(w^{m}) = \sum_{j=1}^{J} g'_j(w^{m} \cdot \xi^j) \cdot \xi^j = \sum_{j=1}^{J} g_{mj}'(w^{m} \cdot \xi^mj) \cdot \xi^mj = \sum_{j=1}^{J} p_{m,j}^{j}
\]
\[ \Delta w^{m,j+1} = -\eta_m p^{m,j} + \tau_{m,j} \Delta w^{m,j-1}, \quad j = 1, \ldots, J; \quad m = 0,1, \ldots \tag{14} \]

Select the initial \( \eta_0 \in (0,1) \) and a positive constant \( N \), we determine \( \eta_m \) in Reference [20]:

\[ \frac{1}{\eta_{m+1}} = \frac{1}{\eta_m} + N, \quad m = 0,1, \ldots \tag{15} \]

Similarly, as in Reference [13], the momentum coefficient is chosen like this

\[ \tau_{m,j} = \begin{cases} \eta_m^2 \| p^{m,j} \| & \text{if } \| \Delta w^{m,j} \| \neq 0 \\ \| \Delta w^{m,j-1} \| & j = 1, \ldots, J; \quad m = 0,1, \ldots \end{cases} \tag{16} \]

Where \( \| \| \) is the Euclidean norm.

3. Main results

Assumption (A1). \( |g(t)|, |g'(t)|, |g^*(t)| \) are uniformly bounded for \( t \in R \).

The Sigmoid function we often use satisfies this assumption.

Assumption (A2). The number of elements of the following stationary point set is finite

\[ \Omega = \{ w | E_w(w) = 0 \} \tag{17} \]

Theorem 3.1 If Assumption (A1) is valid, then there exists positive numbers \( \tilde{N} \) and \( \tilde{\eta} \), such that for any \( N \geq \tilde{N} \) and \( 0 < \eta_0 \leq \min \{ 1, \tilde{\eta} \} \), such that the iteration (14) is weakly convergent:

\[ \lim_{k \to \infty} E_w(w^k) = 0. \tag{18} \]

Theorem 3.2 If Assumption (A1), (A2) are valid, then there exists positive numbers \( \tilde{N} \) and \( \tilde{\eta} \), such that for any \( N \geq \tilde{N} \) and \( 0 < \eta_0 \leq \min \{ 1, \tilde{\eta} \} \), such that the iteration (14) is strongly convergent, that is, the weight \( w \) will converge to a critical value \( w^* \):

\[ \lim_{k \to \infty} w^k = w^*, \quad E_w(w^*) = 0. \tag{19} \]

4. Proofs

Lemma 4.1 the learning rate sequence \( \{ \eta_m \} \) generated by (15) satisfies:

\[ (i) \quad \eta_m \leq \frac{1}{Nm}, \quad m = 1,2, \ldots, \tag{20} \]
This can be simply proved by induction, here we omit. Next let's make a modification to $E(w^{m+1}) - E(w^m)$, from (5),

$$E(w^{(m+1)}) - E(w^m) = \sum_{j=1}^{J} g_j(w^{(m+1)} \cdot \xi^j) - \sum_{j=1}^{J} g_j(w^m \cdot \xi^j)$$

$$= \sum_{j=1}^{J} \left[ g_j(w^{(m+1)} \cdot \xi^j) - g_j(w^m \cdot \xi^j) \right].$$

Using Taylor's formula to we expand $g_j(w^{(m+1)} \cdot \xi^j)$ at $w^m \cdot \xi^j$ as

$$g_j(w^{(m+1)} \cdot \xi^j) = g_j(w^m \cdot \xi^j) + g_j'(w^m \cdot \xi^j)(w^{(m+1)} - w^m) \cdot \xi^j$$

$$+ \frac{1}{2} g_j''(t_{m,j}) \left[(w^{(m+1)} - w^m) \cdot \xi^j\right]^2$$

$$= g_j(w^m \cdot \xi^j) + g_j'(w^m \cdot \xi^j)(w^{(m+1)} - w^m) \cdot \xi^j + \rho_{m,j}$$

Where $t_{m,j}$ is between $w^m \cdot \xi^j$ and $w^{(m+1)} \cdot \xi^j$, and $\rho_{m,j} = 1/2 g_j''(t_{m,j}) \left[(w^{(m+1)} - w^m) \cdot \xi^j\right]^2.$

Then

$$E(w^{(m+1)}) - E(w^m) = \sum_{j=1}^{J} g_j'(w^m \cdot \xi^j) (w^{(m+1)} - w^m) \cdot \xi^j + \sum_{j=1}^{J} \rho_{m,j}$$

$$= \sum_{j=1}^{J} g_j'(w^m \cdot \xi^m) (w^{(m+1)} - w^m) \cdot \xi^m + \sum_{j=1}^{J} \rho_{m,j}$$

Notice that

$$w^{(m+1)} - w^m = \sum_{j=1}^{J} \Delta w^{m+j} = \sum_{j=1}^{J} (-\eta_m p^{m,j} + \tau_{m,j} \Delta w^{m+j-1}),$$

And

$$p^{m,j} = p^{m,j} + (g_{m,j}'(w^{m+j-1} \cdot \xi^m) - g_{m,j}'(w^m \cdot \xi^m)) \cdot \xi^m.$$
\[ E\left( w^{(m+1)j} \right) - E\left( w^{mj} \right) = \sum_{j=1}^{J} g_{mj} \left( w^{mj} \cdot \xi^{mj} \right) + \sum_{s=1}^{S} \tau_{ms} \Delta w^{mj+1} - \eta_{m} \sum_{j=1}^{J} p^{mj} \]

\[ -\eta_{m} \sum_{j=1}^{J} \left( g_{ms} \left( w^{mj+1} \right) \cdot \xi^{ms} \right) - g_{ms} \left( w^{mj} \cdot \xi^{mj} \right) \cdot \xi^{mj} \]

\[ + \sum_{j=1}^{J} \rho_{mj} \]  

(28)

Set

\[ r_{mj} = \left[ g_{mj} \left( w^{mj+1} \cdot \xi^{mj} \right) - g_{mj} \left( w^{mj} \cdot \xi^{mj} \right) \right] \cdot \xi^{mj} = p^{mj+1} - p^{mj}, \]  

(29)

Then

\[ E\left( w^{(m+1)j} \right) - E\left( w^{mj} \right) = -\eta_{m} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + \left( \sum_{j=1}^{J} p^{mj} \right) \left( \sum_{j=1}^{J} \tau_{mj} \Delta w^{mj+1} \right) \]

\[ -\eta_{m} \left( \sum_{j=1}^{J} p^{mj} \right) \left( \sum_{j=1}^{J} r_{mj} \right) + \sum_{j=1}^{J} \rho_{mj} \]  

(30)

This equation will play an important role in the monotonicity of the error function. Below we can analyze each item on the right side of the above equation.

Lemma 4.2 If Assumption (A1) is valid, for \( 0 < \eta_{0} \leq 1 \), it holds

(i) \[ \left\| r_{mj} \right\| \leq C_{1} \eta_{m} \sum_{k=1}^{J} \left\| p^{mk} \right\|, \quad j = 2, \cdots, J, \]  

(31)

(ii) \[ \left\| w^{(m+1)j} - w^{mj} \right\| \leq \eta_{m} \left\| \sum_{j=1}^{J} p^{mj} \right\| + C_{2} \eta_{m}^{2} \sum_{j=1}^{J} \left\| p^{mj} \right\|. \]  

(32)

Where \( C_{1}, C_{2} \) are positive constants.

Proof : (i) From (29), using the intermediate value theorem, we have

\[ r_{mj} = \left[ g_{mj} \left( w^{mj+1} \cdot \xi^{mj} \right) - g_{mj} \left( w^{mj} \cdot \xi^{mj} \right) \right] \cdot \xi^{mj} \]

\[ = \left[ g_{mj} \left( y_{mj} \right) \left( w^{mj+1} - w^{mj} \right) \xi^{mj} \right] \cdot \xi^{mj}, \]  

(33)
Where \( y_{m,j} \) is between \( w^{m_j+1-j}, \zeta_{m_j} \) and \( w^m, \zeta_{m_j} \).

From Assumption (A1) it is easy to see, there exists \( C'_1 > 0 \), such that

\[
\| r_{m,j} \| \leq C'_1 \left\| w^{m_j+1-j} - w^m \right\|
=
C'_1 \left\| \sum_{k=1}^{j-1} \Delta w^{m_j+k} \right\|
=
C'_1 \left\| \sum_{k=1}^{j-1} \left( -\eta_m p^{m,k} + \tau_{m,k} \Delta w^{m_j+k-1} \right) \right\|.
\]

(34)

\[
= C'_1 \left\| -\eta_m \sum_{k=1}^{j-1} p^{m,k} - \eta_m \sum_{k=1}^{j-1} r_{m,k} + \sum_{k=1}^{j-1} \tau_{m,k} \Delta w^{m_j+k-1} \right\|
\leq C'_1 \sum_{k=1}^{j-1} \left( \eta_m \left\| p^{m,k} \right\| + \eta_m \left\| r_{m,k} \right\| + \tau_{m,k} \left\| \Delta w^{m_j+k-1} \right\| \right)
\]

Combine (16) with (20), we have

\[
\| r_{m,j} \| \leq C'_1 \sum_{k=1}^{j-1} \left( (\eta_m + \eta_m^2) \left\| p^{m,k} \right\| + \eta_m \left\| r_{m,k} \right\| \right)
= C'_1 \eta_m \sum_{k=1}^{j-1} \left( 1 + \eta_m \right) \left\| p^{m,k} \right\| + \left\| r_{m,k} \right\|
\leq 2C'_1 \eta_m \sum_{k=1}^{j-1} \left( \left\| p^{m,k} \right\| + \left\| r_{m,k} \right\| \right)
\]

(35)

Next we use induction:

When \( j = 2, r_{m,1} = 0 \), i.e. \( C'_2 = 2C'_1 \), then

\[
\| r_{m,2} \| \leq C'_2 \eta_m \sum_{k=1}^{j-1} \left\| p^{m,k} \right\|.
\]

(36)

When \( 2 \leq j \leq i-1, 3 \leq i \leq J \), suppose

\[
\| r_{m,j} \| \leq C'_j \eta_m \sum_{k=1}^{j-1} \left\| p^{m,k} \right\|,
\]

(37)

Where \( C'_j > 0 \), then
where $\hat{C}_{i-1} = \max \left\{ C_i^{j'} \right\}_{j=1}^{i-1}$.

For all of the $C_j'$ above, set $C_i = \max \left\{ C_i^{j'} \right\}_{j=1}^{i}$, then $\| r_{m,j} \| \leq C_i \eta_m \sum_{i=1}^{i} \| p^{m,j} \|$, $j = 2, \ldots, J$.

(ii) According to Assumptions (A1) and (16), (26), (29), we have

$$
\| w^{(m+1)} - w^m \| = \left\| \sum_{j=1}^{J} r_{m,j} \Delta W^{m+j-1} - \eta_m \sum_{j=1}^{j} p^{m,j} - \eta_m \sum_{j=1}^{j} r_{m,j} \right\|
\leq \sum_{j=1}^{j} \eta_m \| p^{m,j} \| + \eta_m \left\| \sum_{j=1}^{j} r_{m,j} \right\|
\leq \sum_{j=1}^{j} \eta_m \| p^{m,j} \| + \eta_m \left\| \sum_{j=1}^{j} r_{m,j} \right\|
\leq \left[ 1 + (J-1)C_i \right] \eta_m \sum_{j=1}^{j} \| p^{m,j} \| + \eta_m \left\| \sum_{j=1}^{j} r_{m,j} \right\|
\leq C_2 \eta_m \sum_{j=1}^{j} \| p^{m,j} \| + \eta_m \left\| \sum_{j=1}^{j} r_{m,j} \right\|
$$

Where $C_2 = 1 + (J-1)C_i$.

Lemma 4.3 If Assumption (A1) is valid, for $0 < \eta_0 \leq 1$, it holds

(i) $\left\| \left( \sum_{j=1}^{j} p^{m,j} \right) \left( \sum_{j=1}^{j} r_{m,j} \Delta W^{m+j-1} \right) \right\| \leq \eta_m^2 \left( \sum_{j=1}^{j} \| p^{m,j} \| \right)^2$, (40)

(ii) $\left\| \eta_m \left( \sum_{j=1}^{j} p^{m,j} \right) \left( \sum_{j=1}^{j} r_{m,j} \right) \right\| \leq C_3 \eta_m^2 \left( \sum_{j=1}^{j} \| p^{m,j} \| \right)^2$, (41)
(iii) \[ \sum_{j=1}^{l} \rho_{m,j} \leq C_2 \eta m^2 \left( \sum_{j=1}^{l} \| p^{m,j} \| \right)^2. \] (42)

Proof: (i) Form (16), \[ \sum_{j=1}^{l} \tau_{m,j} \Delta W^{m+j-1} \leq \eta m^2 \sum_{j=1}^{l} \| p^{m,j} \|, \] then

\[ \left( \sum_{j=1}^{l} p^{m,j} \right) \left( \sum_{j=1}^{l} \tau_{m,j} \Delta W^{m+j-1} \right) \leq \sum_{j=1}^{l} p^{m,j} \cdot \eta m^2 \sum_{j=1}^{l} \| p^{m,j} \| \]
\[ \leq \eta m^2 \left( \sum_{j=1}^{l} \| p^{m,j} \| \right)^2. \] (43)

(ii) From (31)

\[ \eta m \left( \sum_{j=1}^{l} p^{m,j} \right) \left( \sum_{j=1}^{l} r_{m,j} \right) \leq \eta m \left( \sum_{j=1}^{l} \| p^{m,j} \| \right) \cdot \left( \sum_{j=1}^{l} \| r_{m,j} \| \right) \]
\[ \leq \eta m \left( \sum_{j=1}^{l} \| p^{m,j} \| \right) \left( \sum_{j=1}^{l} C \eta m \sum_{k=1}^{l+1} \| p^{m,k} \| \right) \]
\[ \leq JC \eta m^2 \left( \sum_{j=1}^{l} \| p^{m,j} \| \right)^2 \]
\[ =JC \eta m^2 \left( \sum_{j=1}^{l} \| p^{m,j} \| \right)^2. \] (44)

Where \( C_3 = JC_1 \).

(iii) From Assumption (A1) and \( \rho_{m,j} = \frac{1}{2} g^* (t_{m,j}) \left[ (w^{(m+1)j} - w^{mj}) \cdot \xi_j \right]^2 \), there exists \( C_3 > 0 \), such that

\[ \left| \sum_{j=1}^{l} \rho_{m,j} \right| = \left| \sum_{j=1}^{l} \frac{1}{2} g^* (t_{m,j}) \left[ (w^{(m+1)j} - w^{mj}) \cdot \xi_j \right]^2 \right| \]
\[ \leq C_3 \sum_{j=1}^{l} \| w^{(m+1)j} - w^{mj} \|^2, \] (45)

Combine (32), we have
Where $C_4 = JC_s (1 + C_s)^2$.

Lemma 4.4 If Assumption (A1) is valid, for $0 < \eta_0 \leq 1$, it holds

$$E\left( w^{(m+1)J} \right) - E\left( w^{mj} \right) \leq -\eta_m \left[ \sum_{j=1}^{J} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + C_5 \eta_m^2 \sum_{j=1}^{J} \left\| p^{mj} \right\|^2 \right].$$

Proof: From (30) and Lemma 4.3, we have

$$E\left( w^{(m+1)J} \right) - E\left( w^{mj} \right) = -\eta_m \left[ \sum_{j=1}^{J} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + \left( \sum_{j=1}^{J} p^{mj} \right) \left( \sum_{j=1}^{J} \tau_{mj} \Delta w^{mj+1} \right) \right]$$

$$= -\eta_m \left[ \sum_{j=1}^{J} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + \left( \sum_{j=1}^{J} \tau_{mj} \Delta w^{mj+1} \right) \right]$$

$$\leq -\eta_m \left[ \sum_{j=1}^{J} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + \eta_m^2 \left( 1 + C_3 + C_s \right) \left( \sum_{j=1}^{J} \left\| p^{mj} \right\|^2 \right) \right],$$

$$\leq -\eta_m \left[ \sum_{j=1}^{J} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + \eta_m^2 J (1 + C_3 + C_s) \sum_{j=1}^{J} \left\| p^{mj} \right\|^2 \right]$$

$$= -\eta_m \left[ \sum_{j=1}^{J} \left\| \sum_{j=1}^{J} p^{mj} \right\|^2 + C_5 \eta_m^2 \sum_{j=1}^{J} \left\| p^{mj} \right\|^2 \right].$$

Where $C_s = 1 + C_3 + C_s$.

Lemma 4.5 If Assumption (A1), (A2) are valid, then there exists positive numbers $\tilde{N}$ and $\tilde{\eta}$, such that for any $N \geq \tilde{N}$ and $0 < \eta_0 \leq \min \{ 1, \tilde{\eta} \}$, it holds

$$E\left( w^{mj} \right) \geq E\left( w^{(m+1)J} \right), \ m = 0, 1, \ldots.$$
When \( m = 0 \), take \( \eta_0 > 0 \), such that

\[
\eta_0 \left\| \sum_{j=1}^{J} p_{0,j} \right\|^2 \geq C_5 \eta_0^2 \sum_{j=1}^{J} \left\| p_{0,j} \right\|^2 .
\]  

(50)

For \( m \geq 0 \), suppose

\[
\eta_m \left\| \sum_{j=1}^{J} p_{m,j} \right\|^2 \geq C_5 \eta_m^2 \sum_{j=1}^{J} \left\| p_{m,j} \right\|^2 .
\]  

(51)

Next we only need to prove

\[
\eta_{m+1} \left\| \sum_{j=1}^{J} p_{m+1,j} \right\|^2 \geq C_5 \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m+1,j} \right\|^2 .
\]  

(52)

Notice that \( p_{m+1,j} = g_{(m+1)j} \left( w_{(m+1)j} \cdot \zeta_{(m+1)j} \right) \cdot \zeta_{(m+1)j} \), where \( \{k_1, k_2, \ldots, k_J\} \) is a permutation of \( \{1, 2, \ldots, J\} \). Using the middle value theorem, then

\[
p_{m+1,j} = g_{mk_j} \left( w_{mk_j} \cdot \zeta_{mk_j} \right) \cdot \zeta_{mk_j} + \left[ g_{mk_j} \cdot (s_{mk_j}) \cdot (w_{mk_j} - w_{mj}) \cdot \zeta_{mk_j} \right] \cdot \zeta_{mk_j}
\]  

(53)

Where \( s_{mk_j} \) is between \( w_{mk_j} \cdot \zeta_{mk_j} \) and \( w_{mj} \cdot \zeta_{mk_j} \).

From Assumption (A1) and Lemma 4.2, there exists positive constant \( \tilde{C}_1 \), such that

\[
\left\| p_{m+1,j} \right\| \leq \left\| p_{mk_j} \right\| + \tilde{C}_1 \left\| w_{mk_j} - w_{mj} \right\|
\]  

(54)

\[
\leq \left\| p_{mk_j} \right\| + \tilde{C}_1 \eta_m \left\| \sum_{j=1}^{J} p_{m,j} \right\| + \tilde{C}_1 C_2 \eta_m^2 \sum_{j=1}^{J} \left\| p_{m,j} \right\| .
\]

Take the square of both sides of the above inequality to get
\[
\left\| p^{m+1,j} \right\|^2 \leq \left\| p^{m,j} \right\|^2 + C_2^2 \eta_m^2 \left( \sum_{j=1}^J p^{m,j} \right)^2 + C_1^2 C_2^2 \eta_m^4 \left( \sum_{j=1}^J \left\| p^{m,j} \right\|^2 \right)
\]

\[
+2 \left[ \tilde{C}_1 \eta_m \left\| p^{m,j} \right\|^2 \right] + \tilde{C}_1 C_2 \eta_m^2 \left\| p^{m,j} \right\| \sum_{j=1}^J \left\| p^{m,j} \right\| + \tilde{C}_1 C_2 \eta_m^2 \left\| p^{m,j} \right\| \sum_{j=1}^J \left\| p^{m,j} \right\|
\]

\[
+\tilde{C}_1^2 C_2 \eta_m^3 \left( \sum_{j=1}^J \left\| p^{m,j} \right\|^2 \right)
\]

(56)

Sum both sides of the above inequality over \( j = 1, \ldots, J \), we get

\[
\sum_{j=1}^J \left\| p^{m+1,j} \right\|^2 \leq \left\| p^{m,j} \right\|^2 + J\tilde{C}_1^2 \eta_m^2 \left( \sum_{j=1}^J p^{m,j} \right)^2 + J^2 \tilde{C}_1^2 C_2 \eta_m^4 \left( \sum_{j=1}^J \left\| p^{m,j} \right\|^2 \right)
\]

\[
+\tilde{C}_1 \eta_m \sum_{j=1}^J \left\| p^{m,j} \right\|^2 + J\tilde{C}_1 \eta_m \left\| \sum_{j=1}^J p^{m,j} \right\|^2 + \tilde{C}_1 C_2 \eta_m^2 \left\| \sum_{j=1}^J p^{m,j} \right\|^2
\]

\[
+J^2 \tilde{C}_1 C_2 \eta_m^3 \left( \sum_{j=1}^J \left\| p^{m,j} \right\|^2 \right)
\]

\[
+J^2 \tilde{C}_1 \eta_m \left( \sum_{j=1}^J \left\| p^{m,j} \right\|^2 \right)
\]

(57)

\[
= \sum_{j=1}^J \left\| p^{m,j} \right\|^2 + \left( \left( J\tilde{C}_1 \eta_m + J\tilde{C}_1 + J\tilde{C}_1^2 C_2 \eta_m^2 \right) \eta_m \sum_{j=1}^J \left\| p^{m,j} \right\|^2 \right)
\]

\[
+\left( J^2 \tilde{C}_1 \eta_m^3 \right) \eta_m \sum_{j=1}^J \left\| p^{m,j} \right\|^2
\]

Set \( C_6 = \max \left\{ J\tilde{C}_1^2 + J\tilde{C}_1 + J\tilde{C}_1^2 \right\} \), then

\[
\sum_{j=1}^J \left\| p^{m+1,j} \right\|^2 \leq \sum_{j=1}^J \left\| p^{m,j} \right\|^2 + C_6 \eta_m \sum_{j=1}^J \left\| p^{m,j} \right\|^2
\]

\[
+C_6 \eta_m \sum_{j=1}^J \left\| p^{m,j} \right\|^2
\]

(58)

Using (54) again, we have

\[
\sum_{j=1}^J \left\| p^{m+1,j} \right\|^2 = \sum_{j=1}^J \left\| p^{m,j} \right\|^2 + \sum_{j=1}^J \left[ g_{mk}^n \left( s_{m,j} \right) \cdot \left( w^{(m+1)j} - w_{m} \right) \cdot \xi_{mk} \right] \cdot \xi_{mk}^n.
\]

(59)

Similarly from Assumption (A1)
\[
\left\| \sum_{j=1}^{f} p_{m+1,j} \right\| \geq \left\| \sum_{j=1}^{f} p_{m,j} \right\| - J \tilde{C}_1 \left\| w^{(m+1)} \right\| - w^{mj} \\
\geq \left\| \sum_{j=1}^{f} p_{m,j} \right\| - J \tilde{C}_1 \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\| - J \tilde{C}_1 \eta_n^2 \sum_{j=1}^{f} \left\| p_{m,j} \right\| 
\]

(60)

Set \( x = \left\| \sum_{j=1}^{f} p^{m+1,j} \right\| \), \( y = \left\| \sum_{j=1}^{f} p^{m,j} \right\| \), \( z = J \tilde{C}_1 \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\| + J \tilde{C}_1 \eta_n^2 \sum_{j=1}^{f} \left\| p_{m,j} \right\| \), for non-negative numbers \( x, y, z \), if \( x \geq y - z \), then \( x^2 \geq y^2 - 2yz \), so for(60), we have

\[
\left\| \sum_{j=1}^{f} p^{m+1,j} \right\| \geq \left\| \sum_{j=1}^{f} p^{m,j} \right\| - 2 \left\| \sum_{j=1}^{f} p_{m,j} \right\| \left( J \tilde{C}_1 \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\| + J \tilde{C}_1 \eta_n^2 \sum_{j=1}^{f} \left\| p_{m,j} \right\| \right) \\
\geq \left\| \sum_{j=1}^{f} p^{m,j} \right\| - 2 J \tilde{C}_1 \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\| - J \tilde{C}_1 \eta_n^2 \sum_{j=1}^{f} \left\| p_{m,j} \right\| \left( \left\| \sum_{j=1}^{f} p_{m,j} \right\| + J \sum_{j=1}^{f} \left\| p_{m,j} \right\|^2 \right) \\
= \left\| \sum_{j=1}^{f} p^{m,j} \right\| - (2 J \tilde{C}_1 + J \tilde{C}_1 \eta_n) \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\| - J^2 \tilde{C}_1 \eta_n^3 \sum_{j=1}^{f} \left\| p_{m,j} \right\|^2 
\]

(61)

Set \( C_7 = \max \{ 2 J \tilde{C}_1 + J \tilde{C}_1 \eta_n, J^2 \tilde{C}_1 \eta_n \} \), then

\[
\left\| \sum_{j=1}^{f} p^{m+1,j} \right\| \geq \left\| \sum_{j=1}^{f} p^{m,j} \right\| - C_7 \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\| - C_7 \eta_n^3 \sum_{j=1}^{f} \left\| p_{m,j} \right\|^2 
\]

(62)

Combine (15) with (62)

\[
\frac{1}{\eta_{m+1}} \left\| \sum_{j=1}^{f} p^{m+1,j} \right\| \geq \frac{1}{\eta_m} \left\| \sum_{j=1}^{f} p^{m,j} \right\|^2 + N \left\| \sum_{j=1}^{f} p^{m,j} \right\|^2 - C_7 \left\| \sum_{j=1}^{f} p_{m,j} \right\|^2 - C_7 \eta_n \left\| \sum_{j=1}^{f} p_{m,j} \right\|^2 \\
- C_7 \eta_n \sum_{j=1}^{f} \left\| p_{m,j} \right\|^2 - C_7 \eta_n^2 \sum_{j=1}^{f} \left\| p_{m,j} \right\|^2 
\]

(63)

Multiply both sides of the above inequality by \( \eta_{m+1}^2 \)

\[
\left\| \sum_{j=1}^{f} p^{m+1,j} \right\| \geq \eta_{m+1}^2 \left\| \sum_{j=1}^{f} p^{m,j} \right\|^2 + N \eta_{m+1}^2 \left\| \sum_{j=1}^{f} p^{m,j} \right\|^2 - C_7 \eta_n^2 \eta_{m+1} \left\| \sum_{j=1}^{f} p_{m,j} \right\|^2 \\
- C_7 \eta_n \eta_{m+1} \left\| \sum_{j=1}^{f} p_{m,j} \right\|^2 - C_7 \eta_n \eta_{m+1}^2 \left\| \sum_{j=1}^{f} p_{m,j} \right\|^2 - C_7 \eta_n^2 \eta_{m+1}^2 \sum_{j=1}^{f} \left\| p_{m,j} \right\|^2 
\]

(64)

Combine (52) with (58)
Now we only need to prove (67)

\[ N\eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m+1,j}^{m} \right)^2 \geq C_7 N\eta_{m}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 + C_7 N\eta_{m}^2 \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 + C_7 C_6 \eta_{m} \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \]

In order to ensure that (58) holds true, as long as \( \frac{N}{6}\eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \) is greater than or equal to each term on the right side of the inequality.

Item 1: \( \frac{N}{6}\eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \geq C_7 \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \) needs \( N \geq 6C_7 \);

Item 2: \( \frac{N}{6}\eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \geq C_7 N\eta_{m} \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \) needs \( \eta_{m} \leq \frac{1}{6C_7} \);

Item 3: \( \frac{N}{6}\eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \geq C_7 \eta_{m} \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \), that is \( \frac{N}{6} \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \geq C_7 \eta_{m} \eta_{m+1}^2 \left( \sum_{j=1}^{j} p_{m,j}^{m} \right)^2 \), needs \( N \geq \frac{6C_7}{C_5} \).
Item 4: $\frac{N}{6} \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \geq C_n N \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2$, that is

$$\frac{N}{6} C_s \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \geq C_n N \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2$$ Needs $\eta_m \leq \frac{C_n}{6C_s}$;

Item 5: $\frac{N}{6} \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \geq C_s C_d \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2$ needs $\eta_m \leq \frac{N}{6C_s C_d}$;

Item 6: $\frac{N}{6} \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \geq C_s C_d \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2$, that is

$$\frac{N}{6} C_s \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \geq C_s C_d \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2$$ Needs $N \geq 6C_6$.

To sum up, let

$$\tilde{N} = \max \left\{ 6C_7, \frac{6C_7}{C_5}, 6C_6 \right\},$$

(68)

$$\tilde{\eta} = \min \left\{ 1, \frac{C_5}{6C_7}, \frac{\tilde{N}}{6C_7}, \frac{6C_7}{6C_s C_6}, \frac{1}{6C_7}, \frac{6C_7}{6C_7}, \frac{\sum_{j=1}^{J} \left\| p_{0,j}^{\eta} \right\|^2}{C_5 \left( \sum_{j=1}^{J} \left\| p_{0,j}^{\eta} \right\|^2 \right)} \right\}.$$  (69)

Then for any $N \geq \tilde{N}$ and $0 < \eta_0 \leq \min \{1, \tilde{\eta} \}$, it holds $E\left( w^{mJ} \right) \geq E\left( w^{(m+1)J} \right)$, $m = 0, 1, \ldots$.

Lemma 4.6 If Assumption (A1) is valid, and $\tilde{N}, \tilde{\eta}$ satisfy (68), (69), then for any $N \geq \tilde{N}$ and $0 < \eta_0 \leq \min \{1, \tilde{\eta} \}$, it holds

$$\sum_{m=1}^{\infty} \frac{1}{m} \left\| \sum_{j=1}^{J} p_{m,j}^{\eta} \right\|^2 < \infty.$$  (70)

Proof: From (47) we have $E\left( w^{mJ} \right) - E\left( w^{(m+1)J} \right) \geq \eta_m \left\| \sum_{j=1}^{J} p_{m,j}^{\eta} \right\|^2 - C_s C_d \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2$, sum both sides of the above inequality over $m = 1, 2, \cdots, M$, we get

$$E\left( w^{J} \right) - E\left( w^{(M+1)J} \right) \geq \sum_{m=1}^{M} \left( \eta_m \left\| \sum_{j=1}^{J} p_{m,j}^{\eta} \right\|^2 - C_s C_d \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \right) + \sum_{m=M+1}^{M} \left( \eta_m \left\| \sum_{j=1}^{J} p_{m,j}^{\eta} \right\|^2 - C_s C_d \eta_{m+1}^2 \sum_{j=1}^{J} \left\| p_{m,j}^{\eta} \right\|^2 \right)$$

(71)

From Lemma 4.5, let $M \rightarrow \infty$, it holds
According to Assumption (A1), there exists positive constant $C_{S}$, such that
\[
\sum_{j=1}^{J_{m}} \left\| p_{m,j} \right\|^2 \leq C_{S},
\]
and from (20), it holds
\[
\sum_{m=1}^{\infty} \left( C_{S} \eta_{m}^{2} \sum_{j=1}^{J_{m}} \left\| p_{m,j} \right\|^2 \right) \leq \sum_{m=1}^{\infty} \frac{C_{S} C_{8}}{N^{2} m^{2}} < \infty.
\]
Then combine (72) with (73)
\[
\sum_{m=1}^{\infty} \eta_{m} \left( \sum_{j=1}^{J_{m}} \left\| p_{m,j} \right\|^2 \right) < \infty.
\]
Review (21, $\eta_{m} = \frac{\eta_{0}}{1 + N\eta_{m} m} = o\left(\frac{1}{m}\right)$, then
\[
\sum_{m=1}^{\infty} \left( \frac{1}{m} \sum_{j=1}^{J_{m}} \left\| p_{m,j} \right\|^2 \right) < \infty.
\]
Lemma 4.7 If $\sum_{i=1}^{\infty} \frac{a_{i}^{2}}{i} < \infty$, all of the $a_{i} \geq 0$ and $|a_{i} - a_{i+1}| \leq \frac{1}{i}$, then $\lim_{i \to \infty} a_{i} = 0$.

Lemma 4.7 Let $f : \mathbb{R}^{n} \to \mathbb{R}$ be continuously differentiable. If the number of elements of set $\Omega = \{ x \, | \, f_{x}(x) = 0 \}$ is finite, and the sequence $\{ x^{k} \}$ satisfies
\[
\lim_{k \to \infty} \left\| x^{k} - x^{k+1} \right\| = 0 ; \lim_{k \to \infty} \left\| f_{x}(x^{k}) \right\| = 0,
\]
Then
\[
\lim_{k \to \infty} x^{k} = x^{*} , \quad f_{x}(x^{*}) = 0.
\]
The above two lemmas can be found in Reference [21], here we omit it.
Proof of Theorem 3.1:
We need to prove $\lim_{k \to \infty} E_{n}(w^{m+j}) = 0, \ j = 0, 1, \cdots, J - 1$, we can firstly prove when
\[
j = 0, \ \lim_{k \to \infty} E_{n}(w^{m}) = 0.
\]
According to Lemma 4.2 and Assumption (A1), there exists positive constant $C_9$, such that
\begin{equation}
\| w^{(m+1)J} - w^{mJ} \| \leq C_9 \eta_m. \tag{79}
\end{equation}

Then for any vector $e \in \mathbb{R}^n$, combine (13) with (54), we have
\begin{equation}
E_n(\{w^{(m+1)J}\}) \cdot e - E_n(\{w^{mJ}\}) \cdot e = \left( \sum_{j=1}^J p^{m+1,j} - \sum_{j=1}^J p^{m,j} \right) \cdot e \\
= \sum_{j=1}^J \left[ g_{mk_j}''(s_{m,j}) \cdot (w^{(m+1)J} - w^{mJ}) \cdot \xi^m_{k_j} \right] \cdot (\xi^m_{k_j} \cdot e) \tag{80}
\end{equation}

From Assumption (A1), there exists positive constant $C_{10}$, such that
\begin{equation}
\left| E_n(\{w^{(m+1)J}\}) \cdot e - E_n(\{w^{mJ}\}) \cdot e \right| \leq C_{10} \| e \| \| w^{(m+1)J} - w^{mJ} \| \tag{81}
\end{equation}

According to (20), (79) and (80), it holds
\begin{equation}
\left| E_n(\{w^{(m+1)J}\}) \cdot e - E_n(\{w^{mJ}\}) \cdot e \right| \leq \left| E_n(\{w^{(m+1)J}\}) \cdot e - E_n(\{w^{mJ}\}) \cdot e \right|
\leq C_9 C_{10} \| e \| \eta_m \\
\leq \frac{C_9 C_{10} \| e \|}{N} \tag{82}
\end{equation}

And
\begin{equation}
\left| E_n(\{w^{mJ}\}) \cdot e \right|^2 = \left| \sum_{j=1}^J p^{m,j} \cdot e \right|^2 \leq \left| \sum_{j=1}^J p^{m,j} \right|^2 \cdot \| e \|^2 \tag{83}
\end{equation}

From Lemma 4.6, it holds
\begin{equation}
\sum_m \left( \frac{1}{m} \left| E_n(\{w^{mJ}\}) \cdot e \right|^2 \right) \leq \| e \|^2 \sum_m \left( \frac{1}{m} \left| \sum_{j=1}^J p^{m,j} \right|^2 \right) < \infty \tag{84}
\end{equation}

From (83) (84) and Lemma 4.7, it holds
\begin{equation}
\left| E_n(\{w^{mJ}\}) \cdot e \right| \to 0, m \to \infty. \tag{85}
\end{equation}

Because of the arbitrariness of $e$, $\lim_{m \to \infty} E_n(\{w^{mJ}\}) = 0$.

For $j = 1, \ldots, J - 1$, we can prove $\lim_{k \to \infty} E_n(\{w^{mJ+j}\}) = 0$ in a similar way.

Proof of Theorem 3.2:
From (14), (16), (20) and Assumption (A1), we can get
\[
\lim_{m \to \infty} \| \Delta w^{m+j} \| = \lim_{m \to \infty} \| w^{m+j} - w^{m+j-1} \| = 0, \quad j = 1, \ldots, J
\]  \quad (86)

That is

\[
\lim_{k \to \infty} \| w^k - w^{k+1} \| = 0 \quad (87)
\]

From Assumption (A1), Assumption (A2) and Theorem 3.1, the conditions of Lemma 4.8 are all satisfied, then it holds

\[
\lim_{k \to \infty} w^k = w^*, \quad E_w(w^*) = 0. \quad (88)
\]

5. Conclusions

This paper mainly studies the convergence of the online gradient method with momentum for the two-layer BP neural network. Choose an appropriate learning step size and select the momentum coefficient in an adaptive manner. Assuming that the activation function \(|g(t)|, |g'(t)|, \text{ and } |g''(t)|\) are uniformly bounded, and the samples are randomly arranged in each iteration cycle, then the algorithm has a weak convergence result. In addition, if the number of the elements of the set \(\Omega = \{ w|E_w(w) = 0 \}\) is finite, the algorithm further has a strong convergence result.

In the follow-up research, we can try to randomly arrange the training samples in each iteration cycle of the training samples, group them, and update the weights once for each group; we can also study convergence of the online gradient method with momentum for multi-layer BP neural network.

References

[1] T.L. Fine, S. Mukherjee, Parameter Convergence and Learning Curves for Neural Networks, J. Neural Computation. 11 (1999) 747-769.
[2] W. Finnoff, Diffusion Approximations for the Constant Learning Rate Backpropagation Algorithm and Resistance to Local Minima, J. Neural Computation. 6 (1994) 285-295.
[3] C. Lin, Projected Gradient Methods for Nonnegative Matrix Factorization, J. Neural Computation. 19 (2014) 2756-2779.
[4] W. Wu, D. Xu, Z.X. Li, Convergence of gradient method for Elman networks, J. Applied Mathematics and Mechanics, English Edition. 29 (2008) 1231-1238.
[5] D. Xu, Z.X. Li, W. Wu, Convergence of approximated gradient method for Elman network, J. Neural Network World. 18 (2008) 171-180.
[6] Z. Luo, On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks, J. Neural Computation. 2 (1991) 226-245.
[7] W. Wu, Y. Xu, Deterministic convergence of an online gradient method for neural networks, J. Journal of Computational & Applied Mathematics. 144 (2002) 335-347.
[8] W. Wu, G. Feng, X. Li, Training Multilayer Perceptrons Via Minimization of Sum of Ridge Functions, J. Advances in Computational Mathematics. 17 (2002) 331-347.
[9] M. Torii, M.T. Hagan, Stability of steepest descent with momentum for quadratic functions, J. IEEE Trans Neural Netw. 13 (2002) 752-756.
[10] B. Amit, K. Eugenius, Steepest descent with momentum for quadratic functions is a version of the conjugate gradient method, J. Neural Networks. 17 (2004) 65–71.
[11] I. Ernest, M. Tony, Improved backpropagation learning in neural networks with windowed momentum, J. International Journal of Neural System. 12 (2002) 303-318.
[12] Y. Xiong, C. Zhang, Convergence of an asynchronous batch gradient method with momentum for Pi-sigma networks, J. Mathematica Applicata. 21 (2008) 207-212.
[13] H.F. Lu, W. Wu, Z.X. Li, Convergence of Online Gradient Method with a Penalty Term for BP
Neural Network with Stochastic Inputs, J. Journal of Mathematical Research & Exposition. 27 (2007) 643-653.

[14] L.Q. Zhang, W. Wu, Online gradient methods with a penalty term for neural networks with large training set, J. Nonlinear Dyn. Sci. Technol. 11 (2004) 53–58.

[15] Z.X. Li, W. Wu, Y. Tian, Convergence of an online gradient method for feedforward neural networks with stochastic inputs, J. Journal of Computational and Applied Mathematics. 163 (2003) 165–176.

[16] W. Wu, N.M. Zhang, Z.X. Li, et al, Convergence of Gradient Method with Momentum for Back-Propagation Neural Networks, J. Computational Mathematics, English Edition. 26 (2008) 613-623.

[17] N.M. Zhang, An online gradient method with momentum for two-layer feedforward neural networks, J. IEEE Transactions on Neural Networks. 212 (2006) 488-498.

[18] H. Shao, G. Zheng, Boundedness and convergence of online gradient method with penalty and momentum, J. Neurocomputing. 74 (2011) 765-770.

[19] H. Shao, D. Xu, G. Zheng, et al, Convergence of an online gradient method with inner-product penalty and adaptive momentum, J. Neurocomputing. 77 (2012) 243-252.

[20] W. Wu, G. Feng, X. Li, Training Multilayer Perceptrons Via Minimization of Sum of Ridge Functions, J. Advances in Computational Mathematics. 17 (2002) 331-347.

[21] Y.X. Yuan, W.Y. Sun, Optimization Theory and Methods, Science Press, Beijing, 2001.