Convergence Analysis of Multilayer BP Neural Network with Momentum Term

Xiang Xu*, Gang Xie
East China University of Science and Technology, Shanghai, China

*Corresponding author: xxx1227@foxmail.com

Abstract. In this paper, we analyze the convergence of a back-propagation (BP) neural network with momentum term containing multiple hidden layers. When the learning rate is constant and the momentum coefficient is adaptively changed under certain conditions, we give both the weak and strong convergence results of the algorithm, and give corresponding theoretical proofs for both convergence results.

1. Introduction
The BP algorithm is a widely used algorithm for the training of neural networks. The convergence analysis of the BP algorithm has already been studied [1,2]. It is also a common practice to add momentum to the BP algorithm to accelerate iteration [3-7]. The standard BP algorithm only adjusts weights according to the direction of the error gradient of the current iteration. The momentum BP algorithm also considers the direction of the error gradient of the previous iteration. Adding the momentum term can effectively reduce the trend of oscillation and accelerate the training process. Therefore, the BP algorithm with momentum term now has a tendency to replace the traditional BP algorithm as the new standard BP algorithm.

In the past, we usually chose a constant between 0 and 1 as the momentum coefficient. However, it was found in subsequent simulation experiments that choosing a fixed constant as the momentum coefficient is not a good choice. So now adaptive momentum coefficients are used more often. There are many kinds of choices of adaptive momentum coefficients. In this article, we only choose one of them. Correspondingly, there has been a lot of detailed theoretical analysis of the convergence of these training algorithms [7-11]. But unlike the traditional BP algorithm, the convergence analysis of the momentum BP algorithm is more difficult. Through our research, we find that most of the convergence proofs of the momentum BP algorithm are for simplified cases. For example, the convergence analysis of the momentum BP algorithm without a hidden layer is given in [7,8]; the convergence analysis of the momentum BP algorithm with a single hidden layer is given in [9]; the convergence analysis of the momentum BP algorithm with the error function being a linear function is given in [10]. In [11], the authors gave a convergence analysis of the momentum BP algorithm with the momentum coefficient being a constant.

In this paper, we will generalize the convergence results in [9] from the case with a single hidden layer to the case with double hidden layers. In Section 2, we introduce the momentum BP algorithm and discuss its convergence property. The details of the convergence proof are provided in Section 3.
2. Momentum BP algorithm and Its Convergence

Consider a four-layer neural network. The number of neurons in the input layer, the first hidden layer, the second hidden layer, and the output layer are \( l, m, n, \) and 1, respectively. Let \( \xi^j \in \mathbb{R}^l \ (j = 1, \ldots, J) \) be the training input sample and \( O^j \in \mathbb{R}^l \ (j = 1, \ldots, J) \) be the corresponding output. Let \( U = \{u_{ij}\}_{i \neq j} \) be the weight matrix between the input layer and the first hidden layer. Let \( V = \{v_{ij}\}_{i \neq j} \) be the weight matrix between the first hidden layer and the second hidden layer. Let \( w = \{w_{ij}\}_{i \neq j} \) be the weight vector between the second hidden layer and the output layer.

Let \( g : \mathbb{R} \to \mathbb{R} \) the activation function between the layers. For convenience, we define the following vector function:

\[
G(x) = (g(x_1), g(x_2), \ldots, g(x_n)) \quad \text{for} \quad x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n
\]  

(1)

For any given input \( \xi \in \mathbb{R}^l \), the output at the first hidden layer is \( G(U\xi) \), the output at the second hidden layer is \( G(VG(U\xi)) \), and the final output is:

\[
\zeta = g(w \cdot G(VG(U\xi)))
\]

(2)

In practical applications, we should add bias to the hidden layers and the output layer. In order to simplify the presentation and derivation, we have dropped the bias here.

We use the usual squared error function as follows:

\[
E(w, V, U) := \frac{1}{2} \sum_{j=1}^{J} \left[ O^j - g(w \cdot G(VG(U\xi^j))) \right]^2
\]

\[
= \sum_{j=1}^{J} g_j \left( w \cdot G(VG(U\xi^j)) \right)
\]

(3)

where \( g_j(t) = \frac{1}{2} \left( O^j - g(t) \right)^2 \). \( j = 1, 2, \ldots, J \). The goal of this network training is to find \((w^*, V^*, U^*)\) to make sure:

\[
E(w^*, V^*, U^*) = \min E(w, V, U)
\]

(4)

Calculate the gradients of the error function with respect to the parameters \( w, V, U \):

\[
E_w(w, V, U) = \sum_{j=1}^{J} g_j \left( w \cdot G(VG(U\xi^j)) \right) G(VG(U\xi^j))
\]

(5)

\[
E_v(w, V, U) = \sum_{j=1}^{J} g_j \left( w \cdot G(VG(U\xi^j)) \right) \sum_{i=1}^{n} v_{ij} \cdot g'(v_{ij} \cdot G(U\xi^i)) G(U\xi^i), \ i = 1, 2, \ldots, n
\]

(6)
Given the initial weights $w^0, w^1$, $V^0, V^1$ and $U^0, U^1$ arbitrarily, the weight update method of the momentum BP algorithm is as follows:

$$E_{w^i}(w, V, U) = \sum_{j=1}^{J} \sum_{p=1}^{n} g_j^i \left( w \cdot G \left( V G \left( U \xi^j \right) \right) \right) w_p g' \left( v_p \cdot G \left( U \xi^j \right) \right) v_p g' \left( u_i \cdot \xi^j \right) \xi^j, \quad i = 1, 2, \ldots, m$$  \hspace{1cm} (7)

where $\eta \in (0,1)$ is the learning rate, $\tau_k \cdot \gamma_{k,i},$ and $\gamma_{k,i,j}$ are the momentum coefficients. Denote:

$$\Delta w^{k+1} = w^{k+1} - w^k$$  \hspace{1cm} (9)

$$\Delta v_i^{k+1} = v_i^{k+1} - v_i^k$$  \hspace{1cm} (10)

$$\Delta u_i^{k+1} = u_i^{k+1} - u_i^k$$  \hspace{1cm} (11)

$$p^k = E_{w^i}(w^k, V^k, U^k) = \sum_{j=1}^{J} g_j^i \left( w^k \cdot G \left( V^k G \left( U^k \xi^j \right) \right) \right) G \left( V^k G \left( U^k \xi^j \right) \right)$$  \hspace{1cm} (12)

$$q_i^k = E_{v_i}(w^k, V^k, U^k) = \sum_{j=1}^{J} g_j^i \left( w^k \cdot G \left( V^k G \left( U^k \xi^j \right) \right) \right) w_i^k$$  \hspace{1cm} (13)

$$g' \left( v_i^k \cdot G \left( U^k \xi^j \right) \right) G \left( U^k \xi^j \right) \quad i = 1, 2, \ldots, n$$

$$r_i^k = E_{u_i}(w^k, V^k, U^k) = \sum_{j=1}^{J} \sum_{p=1}^{n} g_j^i \left( w^k \cdot G \left( V^k G \left( U^k \xi^j \right) \right) \right) w_p^k$$  \hspace{1cm} (14)

$$g' \left( v_p^k \cdot G \left( U^k \xi^j \right) \right) v_p^k g' \left( u_i^k \cdot \xi^j \right) \xi^j \quad i = 1, 2, \ldots, m$$

Then (8) can be simplified as:
Similar to [7], the choices of momentum coefficients \(\tau_k\), \(\gamma_{k,i,v}\), and \(\gamma_{k,j,u}\) are as follows:

\[
\begin{align*}
\tau_k &= \begin{cases} 
\frac{\tau_p}{\|\Delta w^k\|} & \text{if } \|\Delta w^k\| \neq 0 \\
0 & \text{else}
\end{cases} \\
\gamma_{k,i,v} &= \begin{cases} 
\frac{\tau_q}{\|\Delta v^k_i\|} & \text{if } \|\Delta v^k_i\| \neq 0 \\
0 & \text{else}
\end{cases} \\
\gamma_{k,j,u} &= \begin{cases} 
\frac{\tau_r}{\|\Delta u^k_j\|} & \text{if } \|\Delta u^k_j\| \neq 0 \\
0 & \text{else}
\end{cases}
\end{align*}
\]

(16)

where \(\tau \in (0,1)\) is the momentum factor, \(\|\|\) is the Euclidian norm.

The following assumptions are given:

(A1) for \(t \in R\), \(|g(t)|, |g'(t)|\) and \(|g''(t)|\) are uniformly bounded

(A2) \(|w^k|(k = 1, 2, \ldots)\) are uniformly bounded

(A3) The following set contains finite elements:

\[
\Omega = \left\{ (w, V, U) \mid E_w \left( w^k, V^k, U^k \right) = 0, E_{v_i} \left( w^k, V^k, U^k \right) = 0, E_{u_j} \left( w^k, V^k, U^k \right) = 0 \right\}
\]

(17)

Remark 1.1: For common activation functions such as Sigmoid functions, assumption (A1) is valid. Condition (A2) is to ensure the weak convergence (19)-(22) in the following theorem, namely, the boundedness guarantees the weak convergence in the learning iteration process. Assumption (A3) requires that the error function has only a limited number of local minimum points. This is to ensure strong convergence (23)-(24) in the following theorem. It is easy to verify from (A1) (A2) that \(\|G(x)\|\) is bounded for any \(x \in R^n\), that \(|p^k|, |q^k_i|(i = 1, 2, \ldots, n)\), \(|r^k_i|(i = 1, 2, \ldots, m)\) are uniformly bounded, and that \(|g_j(t)|, |g'_j(t)|\) and \(|g''_j(t)|(j = 1, 2, \ldots, J)\) are also uniformly bounded for \(t \in R\). These observations will be frequently used later in our proofs.

Theorem 1.1 If (A1) and (A2) are valid, there exist constants \(C^* > 0\) and \(E^* \geq 0\) such that for \(0 < s < 1\), \(\tau = s\eta\) and

\[
\eta < \min \left\{1, \frac{1 - s}{C^* \left( \frac{(2 + s + s^2)^4}{(4 + 3s)(5 + 3s + 2s^2)} \right)} \right\}
\]

(18)

the following results hold for iteration progress (15),

\[
\lim_{k \to \infty} E \left( w^k, V^k, U^k \right) = E^*
\]

(19)
Theorem 1.2 If (A1), (A2) and (A3) are all valid, the iteration progress (15) can converge to a local minimum \((w^*, V^*, U^*)\):

\[
\lim_{k \to \infty} E_n(w^k, V^k, U^k) = 0
\]

\[
\lim_{k \to \infty} E_n(w^k, V^k, U^k) = 0, \quad i = 1, 2, \ldots n
\]

\[
\lim_{k \to \infty} E_n(w^k, V^k, U^k) = 0, \quad i = 1, 2, \ldots m
\]

3. Proof of Theorems

First, we give some definitions to simplify our presentation:

\[
U^{k,j} = G(U^k z^j), V^{k,j} = G(V^k U^{k,j}), w^{k,j} = w^k \cdot V^{k,j}
\]

3.1. Error function expansion

Because of the Taylor’s expansion formula, we can expand \(g_j(w^{k+1,j})\) at \(w^{k,j}\):

\[
g_j(w^{k+1,j}) = g_j(w^{k,j}) + g'_j(w^{k,j})(w^{k+1,j} - w^{k,j}) + \frac{1}{2} g''_j(t_{k,j})(w^{k+1,j} - w^{k,j})^2
\]

where \(t_{k,j}\) is between \(w^{k+1,j}\) and \(w^{k,j}\).

Combining formula (3) and the above formula leads to:

\[
E(w^{k+1}, V^{k+1}, U^{k+1}) = E(w^k, V^k, U^k) + \sum_{j=1}^{J} g'_j(w^{k,j})(w^{k+1,j} - w^{k,j}) + \frac{1}{2} \sum_{j=1}^{J} g''_j(t_{k,j})(w^{k+1,j} - w^{k,j})^2
\]

Obviously, \(w^{k+1,j} - w^{k,j} = V^{k,j} \cdot \Delta w^{k+1} + (V^{k+1,j} - V^{k,j}) \cdot w^{k+1}\).

Then by formula (15),

\[
\sum_{j=1}^{J} g'_j(w^{k,j})(w^{k+1,j} - w^{k,j}) = \sum_{j=1}^{J} g'_j(w^{k,j})V^{k,j}(\tau \Delta w^k - \eta p^k) + \sum_{j=1}^{J} g'_j(w^{k,j})(V^{k+1,j} - V^{k,j}) \cdot w^{k+1}
\]

\[
= -\eta \|p^k\|^2 + \tau \cdot p^k \cdot \Delta w^k + \sum_{j=1}^{J} g'_j(w^{k,j})(V^{k+1,j} - V^{k,j}) \cdot w^{k+1}
\]
Using Taylor’s formula again, there exists \( \hat{t}_{k,i,j} \) between \( v_{i}^{k+1} \cdot U_{k+1,j} \) and \( v_{i}^{k} \cdot U_{k,j} \) such that:

\[
\begin{align*}
g(v_{i}^{k+1} \cdot U_{k+1,j}) & = g(v_{i}^{k} \cdot U_{k,j}) + g'(v_{i}^{k} \cdot U_{k,j})(v_{i}^{k+1} \cdot U_{k+1,j} - v_{i}^{k} \cdot U_{k,j}) \\
& + \frac{1}{2} g''(\hat{t}_{k,i,j})(v_{i}^{k+1} \cdot U_{k+1,j} - v_{i}^{k} \cdot U_{k,j})^2
\end{align*}
\]

(29)

Equally obviously, 
\[ v_{i}^{k+1} \cdot U_{k+1,j} - v_{i}^{k} \cdot U_{k,j} = U_{k,j} \cdot \Delta v_{i}^{k+1} + (U_{k+1,j} - U_{k,j}) \cdot v_{i}^{k+1}. \]

and 
\[ (V_{k+1,j} - V_{k,j}) \cdot w_{k+1} = \sum_{i=1}^{n} \left[ g(v_{i}^{k+1} \cdot U_{k+1,j}) - g(v_{i}^{k} \cdot U_{k,j}) \right] \cdot w_{i}^{k+1}. \]

It can be seen that 
\[ \sum_{j=1}^{J} g'(w_{k,j})(V_{k+1,j} - V_{k,j}) \cdot w_{k+1} = \sum_{j=1}^{J} g'(w_{k,j}) \sum_{i=1}^{n} \left[ g'(v_{i}^{k} \cdot U_{k,j}) U_{k,j} \cdot \Delta v_{i}^{k+1} + 
\begin{align*}
& g'(v_{i}^{k} \cdot U_{k,j})(U_{k+1,j} - U_{k,j}) \cdot v_{i}^{k+1} + \frac{1}{2} g''(\hat{t}_{k,i,j})(v_{i}^{k+1} \cdot U_{k+1,j} - v_{i}^{k} \cdot U_{k,j})^2
\end{align*}
\]

Then, the above formula can be simplified by (15):

\[
\begin{align*}
\sum_{j=1}^{J} g'(w_{k,j})(V_{k+1,j} - V_{k,j}) \cdot w_{k+1} = -\eta \sum_{i=1}^{n} \left\| q_{i}^{k} \right\|^2 + \sum_{i=1}^{n} \gamma_{k,i}, q_{i}^{k} \cdot \Delta v_{i}^{k} \\
& + \sum_{j=1}^{J} \sum_{i=1}^{n} g'(w_{k,j}) w_{k}^{j} g'(v_{i}^{k} \cdot U_{k,j})(U_{k+1,j} - U_{k,j}) \cdot v_{i}^{k+1} \\
& + \frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n} g'(w_{k,j}) w_{k}^{j} g''(\hat{t}_{k,i,j})(v_{i}^{k+1} \cdot U_{k+1,j} - v_{i}^{k} \cdot U_{k,j})^2 \\
& + \sum_{j=1}^{J} g'(w_{k,j})(V_{k+1,j} - V_{k,j}) \cdot \Delta w_{k+1}^{j}
\end{align*}
\]

(30)

Using Taylor’s formula again, there exists \( \hat{s}_{k,i,j} \) between \( u_{i}^{k+1} \cdot \xi_{j} \) and \( u_{i}^{k} \cdot \xi_{j} \) such that:

\[
\begin{align*}
g(u_{i}^{k+1} \cdot \xi_{j}) & = g(u_{i}^{k} \cdot \xi_{j}) + g'(u_{i}^{k} \cdot \xi_{j}) \Delta u_{i}^{k+1} \cdot \xi_{j} + \frac{1}{2} g''(\hat{s}_{k,i,j})(\Delta u_{i}^{k+1} \cdot \xi_{j})^2
\end{align*}
\]

From (15) the deformable formula is:

\[
\begin{align*}
g(u_{i}^{k+1} \cdot \xi_{j}) - g(u_{i}^{k} \cdot \xi_{j}) & = -\eta g'(u_{i}^{k} \cdot \xi_{j}) r_{k}^{j} \cdot \xi_{j} + g'(u_{i}^{k} \cdot \xi_{j}) \gamma_{k,i}, \nabla u_{i}^{k} \cdot \xi_{j} \\
& + \frac{1}{2} g''(\hat{s}_{k,i,j})(\Delta u_{i}^{k+1} \cdot \xi_{j})^2
\end{align*}
\]

(31)

We rewrite the expression in (30) as:

\[ \sum_{j=1}^{J} \sum_{p=1}^{n} g'(w_{p,k,j}) w_{p}^{j} g'(v_{p}^{k} \cdot U_{k,j})(U_{k+1,j} - U_{k,j}) \cdot v_{p}^{k+1} \]
Equally obviously, \( (U^{k+1,j} - U^{k,j}) \cdot v_p^{k+1} = \sum_{i=1}^{m} \left[ g \left( u_i^{k+1} \cdot \xi_j \right) - g \left( u_i^{k} \cdot \xi_j \right) \right] \cdot \xi_j \).

From (31), we can get:

\[
\sum_{j=1}^{n} \sum_{p=1}^{m} g'_j \left( w^{k,j} \right) v_p^k g'' \left( v_p^k \cdot U^{k,j} \right) \left( U^{k+1,j} - U^{k,j} \right) \cdot v_p^{k+1}
= -\eta \sum_{i=1}^{n} \left\| r_i^k \right\|^2 + \sum_{i=1}^{m} \gamma_{k,i,a} r_i^k \cdot \nabla u_i^k
+ \frac{1}{2} \eta \sum_{j=1}^{n} \sum_{p=1}^{m} g'_j \left( w^{k,j} \right) v_p^k g'' \left( v_p^k \cdot U^{k,j} \right) v_p^k g'' \left( \xi_{k,i,j} \right) \left( \Delta u_i^{k+1} \cdot \xi_j \right)^2
+ \sum_{j=1}^{n} \sum_{p=1}^{m} g'_j \left( w^{k,j} \right) v_p^k g'' \left( v_p^k \cdot U^{k,j} \right) \left( U^{k+1,j} - U^{k,j} \right) \cdot \Delta v_p^{k+1}
\]

Finally, combining (27), (28), (30) and (32) gives:

\[
E \left( w^{k+1}, V^{k+1}, U^{k+1} \right) - E \left( w^k, V^k, U^k \right)
= -\eta \left\| p^k \right\|^2 + \tau_k p^k \cdot \Delta w^k + \frac{1}{2} \eta \sum_{j=1}^{n} g'' \left( t_{k,j} \right) \left( w^{k+1,j} - w^{k,j} \right)^2
- \eta \sum_{i=1}^{n} \left\| q_i^k \right\|^2 + \sum_{i=1}^{m} \gamma_{k,i,a} q_i^k \cdot \Delta v_i^k
+ \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{m} g'_j \left( w^{k,j} \right) v_i^k g'' \left( \xi_{k,i,j} \right) \left( v_i^{k+1} \cdot U^{k+1,j} - v_i^k \cdot U^{k,j} \right)^2
+ \sum_{j=1}^{n} g'_j \left( w^{k,j} \right) \left( V^{k+1,j} - V^{k,j} \right) \cdot \Delta w^{k+1}
- \eta \sum_{i=1}^{n} \left\| r_i^k \right\|^2 + \sum_{i=1}^{m} \gamma_{k,i,a} r_i^k \cdot \nabla u_i^k
+ \frac{1}{2} \sum_{j=1}^{n} \sum_{p=1}^{m} \sum_{i=1}^{m} g'_j \left( w^{k,j} \right) v_p^k g'' \left( v_p^k \cdot U^{k,j} \right) v_p^k g'' \left( \xi_{k,i,j} \right) \left( \Delta u_i^{k+1} \cdot \xi_j \right)^2
+ \sum_{j=1}^{n} \sum_{p=1}^{m} g'_j \left( w^{k,j} \right) v_p^k g'' \left( v_p^k \cdot U^{k,j} \right) \left( U^{k+1,j} - U^{k,j} \right) \cdot \Delta v_p^{k+1}
\]

### 3.2. Useful lemmas

Lemma 3.1 If (A1) and (A2) are valid, there exists constant \( C_0 > 0 \) such that:

\[
\left\| U^{k+1,j} - U^{k,j} \right\| \leq C_0 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{j=1}^{m} \left\| r_i^k \right\|^2 \right)^{\frac{1}{2}}
\]
Proof: Let \( \phi_{k,i,j} = g \left( u_{i}^{k+1} \cdot \xi^{j} \right) - g \left( u_{i}^{k} \cdot \xi^{j} \right) \), so \( \| U^{k+1} - U^{k} \| = \left( \sum_{i=1}^{m} \phi_{k,i,j}^{2} \right)^{1/2} \).

From (31), we can get: \( \phi_{k,i,j} = -\eta g' \left( u_{i}^{k} \cdot \xi^{j} \right) r_{i}^{j} \cdot \xi^{j} + g' \left( u_{i}^{k} \cdot \xi^{j} \right) \gamma_{k,i,j} \nabla u_{i}^{k} \cdot \xi^{j} + \frac{1}{2} g'' \left( \delta_{i,j} \right) \left( \Delta u_{i}^{k+1} \cdot \xi^{j} \right)^{2} \).

It is not difficult to find that there exists \( C_{1} > 0 \) such that:

\[
|\phi_{k,i,j}| \leq C_{1} \left( \eta \| r_{i}^{j} \| + \| \gamma_{k,i,j} \| \| \Delta u_{i}^{k} \| + \| \Delta u_{i}^{k+1} \| \right) \leq C_{1} \left( \eta + \tau \right) \| r_{i}^{j} \| + C_{1} \| \Delta u_{i}^{k+1} \| ^{2} \tag{35}
\]

Combining (15) and (16) gives: \( \| \Delta u_{i}^{k+1} \| ^{2} \leq \left( \| \gamma_{k,i,j} \| \Delta u_{i}^{k} \| + \| \eta \| r_{i}^{j} \| + 2 \left( \tau^{2} + \eta^{2} \right) \right) \| r_{i}^{j} \| ^{2} \).

According to assumptions (A1) and (A2), \( \| r_{i}^{j} \| \) is uniformly bounded. So there exists \( C_{2} > 0 \) such that: for any \( i, k \), \( \| r_{i}^{j} \| \leq C_{2} \), and \( \| \Delta u_{i}^{k+1} \| ^{2} \leq 2C_{2} \left( \tau^{2} + \eta^{2} \right) \| r_{i}^{j} \| ^{2} \)

Thus, \( \| \phi_{k,i,j} \| \leq C_{0} \left( \eta + \tau + \eta^{2} + \tau^{2} \right) \| r_{i}^{j} \| ^{2} \), where \( C_{0} = \max \{ C_{1}, 2C_{1}C_{2} \} \).

Namely, \( \| U^{k+1} - U^{k} \| \leq C_{0} \left( \eta + \tau + \eta^{2} + \tau^{2} \right) \left( \sum_{i=1}^{m} \| r_{i}^{j} \| ^{2} \right)^{1/2} \).

Lemma 3.2 If (A1) and (A2) are valid, there exists constant \( C_{3} > 0 \) such that:

\[
\left| v_{i}^{k+1} \cdot U^{k+1} - v_{i}^{k} \cdot U^{k} \right| \leq C_{3} \left( \eta + \tau \right) \| q_{i}^{j} \| + C_{3} \left( \eta + \tau + \eta^{2} + \tau^{2} \right) \left( \sum_{i=1}^{m} \| r_{i}^{j} \| ^{2} \right)^{1/2} \tag{36}
\]

Proof: Obviously, \( \left| v_{i}^{k+1} \cdot U^{k+1} - v_{i}^{k} \cdot U^{k} \right| = \left| v_{i}^{k+1} \cdot U^{k+1} - v_{i}^{k+1} \cdot U^{k} + v_{i}^{k+1} \cdot U^{k} - v_{i}^{k} \cdot U^{k} \right| \leq \left| U^{k+1} \cdot \Delta v_{i}^{k+1} + \left( U^{k+1} - U^{k} \right) \cdot v_{i}^{k+1} \right| \).

So there exists \( C_{4} > 0 \) such that: \( \left| U^{k+1} \cdot \Delta v_{i}^{k+1} \right| \leq C_{4} \left| \Delta v_{i}^{k+1} \right| \leq C_{3} \left( \eta + \tau \right) \| q_{i}^{j} \| \)

Combining (A2) and (34), we can know that there exists \( C_{4} > 0 \) such that:

\[
\left| U^{k+1} - U^{k} \right| \cdot v_{i}^{k+1} \leq C_{4} \left| U^{k+1} - U^{k} \right| \leq C_{4} \left( \eta + \tau + \eta^{2} + \tau^{2} \right) \left( \sum_{i=1}^{m} \| r_{i}^{j} \| ^{2} \right)^{1/2}
\]

Thus, \( \left| v_{i}^{k+1} \cdot U^{k+1} - v_{i}^{k} \cdot U^{k} \right| \leq C_{5} \left( \eta + \tau \right) \| q_{i}^{j} \| + C_{5} \left( \eta + \tau + \eta^{2} + \tau^{2} \right) \left( \sum_{i=1}^{m} \| r_{i}^{j} \| ^{2} \right)^{1/2} \)

where \( C_{5} = \max \{ C_{3}, C_{4}C_{0} \} \).

Lemma 3.3 If (A1) and (A2) are valid, there exists constant \( C_{9} > 0 \) such that:
\[ \| V^{k+1,j} - V^{k,j} \| \leq C_9 \left( \eta + \tau + (\eta + \tau)^2 \right) \left( \sum_{j=1}^{n} \| q_i^r \|^2 \right)^{1/2} + C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{i=1}^{m} \| r_i^q \|^2 \right)^{1/2} \]
\[ + C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \sum_{i=1}^{m} \| r_i^q \|^2 \] (37)

Proof: Let \( \psi_{k,j} = g \left( v_i^{k+1} \cdot U^{k+1} \right) - g \left( v_i^k \cdot U^k \right) \), so \( \| \psi_{k+1,j} - \psi_{k,j} \| = \left( \sum_{j=1}^{n} \psi_{k,j}^2 \right)^{1/2} \).

From (29), we can get: \( \psi_{k,j} = g' \left( v_i^k \cdot U^k \right) \left( v_i^{k+1} \cdot U^{k+1} - v_i^k \cdot U^k \right) + \frac{1}{2} g''(v_i^k \cdot U^k) (v_i^{k+1} \cdot U^{k+1} - v_i^k \cdot U^k)^2 \)\). Then from (A1), we can know that there exists \( C_6 > 0 \) such that:
\[ \| \psi_{k,j} \| \leq C_6 \| v_i^{k+1} \cdot U^{k+1} - v_i^k \cdot U^k \| + C_6 \| v_i^{k+1} \cdot U^{k+1} - v_i^k \cdot U^k \|^2 \]

Adding (36) gives:
\[ \| \psi_{k,j} \| \leq C_6 C_5 \left( \eta + \tau \right) \| q_i^r \|^2 + 2C_6 C_5 \left( \eta + \tau \right)^2 \| q_i^r \|^2 + C_6 C_5 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{j=1}^{m} \| r_i^q \|^2 \right)^{1/2} \]
\[ + 2C_6 C_5 \left( \eta + \tau + \eta^2 + \tau^2 \right)^2 \sum_{i=1}^{m} \| r_i^q \|^2 \] (38)

According to assumptions (A1) and (A2), \( \| q_i^r \|^2 \) is uniformly bounded. So there exists \( C_7 > 0 \) such that: for any \( i, k \), \( \| q_i^r \| \leq C_7 \). Then the above formula can be simplified as:
\[ \| \psi_{k,j} \| \leq C_8 \left( \eta + \tau \right) \left( \eta + \tau \right) \| q_i^r \|^2 + C_6 C_5 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{j=1}^{m} \| r_i^q \|^2 \right)^{1/2} \]
\[ + 2C_6 C_5 \left( \eta + \tau + \eta^2 + \tau^2 \right)^2 \sum_{i=1}^{m} \| r_i^q \|^2 \]
\[ \psi_{k,j} \leq 4C_8 \left( \eta + \tau \right) \left( \eta + \tau \right) \| q_i^r \|^2 + 4C_6 C_5 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{j=1}^{m} \| r_i^q \|^2 \right)^{1/2} \]
and
\[ + 2 \times 4C_6 C_5 \left( \eta + \tau + \eta^2 + \tau^2 \right)^4 \sum_{j=1}^{m} \| r_i^q \|^2 \]
where \( C_8 = \max \left\{ C_6 C_5, 2C_6 C_5^2 \right\} \).
\[ \left( \sum_{i=1}^{m} \psi_{k,j}^2 \right)^{1/2} \leq C_9 \left( \eta + \tau \right) \left( \eta + \tau \right) \left( \sum_{j=1}^{m} \| q_i^r \|^2 \right)^{1/2} + C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{j=1}^{m} \| r_i^q \|^2 \right)^{1/2} \]
\[ + C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \sum_{i=1}^{m} \| r_i^q \|^2 \]

so,
\[ \left( \sum_{i=1}^{m} \psi_{k,j}^2 \right)^{1/2} \leq C_9 \left( \eta + \tau \right) \left( \eta + \tau \right) \left( \sum_{j=1}^{m} \| q_i^r \|^2 \right)^{1/2} + C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{j=1}^{m} \| r_i^q \|^2 \right)^{1/2} \]
\[ + C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \sum_{i=1}^{m} \| r_i^q \|^2 \]
where $C_0 = \max \{2C_0, 2\sqrt{n}C_0, 2\sqrt{2n}C_0^2\}$.

Lemma 3.4 If (A1) and (A2) are valid, there exists constant $C^* > 0$ such that:

$$
\left\| \frac{1}{2} \sum_{j=1}^{J} g_j^* \left( t_{k,j} \right) \left( w^{k+1,j} - w^{k,j} \right) \right\|^2 \leq C^* \left( \eta + \tau \right)^2 \left\| p^k \right\|^2 + C^* \left( \eta + \tau + \left( \eta + \tau \right)^2 \right) \sum_{i=1}^{n} \left\| q_i^k \right\|^2
$$

(39)

$$
+ C^* \left[ \left( \eta + \tau + \eta^2 + \tau^2 \right)^2 + \left( \eta + \tau + \eta^2 + \tau^2 \right)^4 \right] \sum_{j=1}^{J} \left\| r_j^k \right\|^2
$$

$$
\left\| \frac{1}{2} \sum_{j=1}^{J} \sum_{p=1}^{P} g_j^* \left( w^{k,j} \right) v_p g^* \left( s_{k,j} \right) \left( \Delta u^{k+1,j} \cdot \xi \right)^2 \right\| \leq C^* \left( \eta + \tau \right)^2 \sum_{i=1}^{n} \left\| q_i^k \right\|^2
$$

(40)

$$
+ C^* \left( \eta + \tau \right) \left( \eta + \tau + \left( \eta + \tau \right)^2 \right) \sum_{j=1}^{J} \left\| r_j^k \right\|^2
$$

$$
\left\| \frac{1}{2} \sum_{j=1}^{J} \sum_{p=1}^{P} g_j^* \left( w^{k,j} \right) v_p g^* \left( s_{k,j} \right) \left( \Delta u^{k+1,j} \cdot \xi \right) \right\| \leq C^* \left( \eta + \tau \right)^2 \sum_{i=1}^{n} \left\| q_i^k \right\|^2
$$

(42)

$$
\leq C^* \left( \eta + \tau \right) \left( \eta + \tau + \eta^2 + \tau^2 \right) \sum_{i=1}^{n} \left\| q_i^k \right\|^2 + C^* \left( \eta + \tau \right) \left( \eta + \tau + \eta^2 + \tau^2 \right) \sum_{j=1}^{J} \left\| r_j^k \right\|^2
$$

(43)

Proof: First, we will prove (39). We have known that $w^{k+1,j} - w^{k,j} = V^{k,j} \cdot \Delta w^{k+1} - \left( V^{k+1,j} - V^{k,j} \right) \cdot w^{k+1}$.

So $\left\| w^{k+1,j} - w^{k,j} \right\| \leq \left\| V^{k,j} \cdot \Delta w^{k+1} \right\| + \left\| \left( V^{k+1,j} - V^{k,j} \right) \cdot w^{k+1} \right\|$.

According to assumptions (A1) and (A2), there exists $C_{10} > 0$ such that:

$$
\left\| V^{k,j} \cdot \Delta w^{k+1} \right\| \leq C_{10} \left( \eta + \tau \right) \left\| p^k \right\|
$$

(37)

Considering (37), we can find that there exists $C_{11} > 0$ such that:

$$
\left\| \left( V^{k+1,j} - V^{k,j} \right) \cdot w^{k+1} \right\| \leq C_{11} \left\| V^{k+1,j} - V^{k,j} \right\| \leq C_{11} C_9 \left( \eta + \tau + \left( \eta + \tau \right)^2 \right) \left( \sum_{i=1}^{n} \left\| q_i^k \right\|^2 \right)^{\frac{1}{2}}
$$

$$
+ C_{11} C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right) \left( \sum_{i=1}^{n} \left\| r_i^k \right\|^2 \right)^{\frac{1}{2}} + C_{11} C_9 \left( \eta + \tau + \eta^2 + \tau^2 \right)^2 \sum_{j=1}^{J} \left\| r_j^k \right\|^2
$$
We continue to prove (41). According to (A1), there exists 
where 

The above inequality can be simplified:

Then, the above formula can be simplified:

According to (A1), there exists 

such that:

where 

It follows from (A1), (A2) and Lemma3.2 that there exists 

such that:

Next, we will prove (40). It follows from (A1), (A2) and Lemma3.2 that there exists 

such that:

The above inequality can be simplified:

where 

We continue to prove (41). According to (A1), there exists 

such that:

(44)
It can be easily seen from (15) and (16) that \( \| \Delta w^{k+1} \| \leq (\eta + \tau) \| p^k \| \).

Adding (37) and (44) gives:

\[
\sum_{j=1}^{J} g_j'(w^{k+1/j})(V^{k+1/j} - V^{k+1/j}) \cdot \Delta w^{k+1} \leq JC_{17} C_9 (\eta + \tau) \| p^k \| \left( (\eta + \eta^2 + \tau^2) \left( \sum_{i=1}^{m} \| q_i^k \| \right)^{1/2} \right) \\
+ (\eta + \eta^2 + \tau^2) \left( \sum_{i=1}^{m} \| q_i^k \| \right)^{1/2} + (\eta + \eta^2 + \tau^2) \sum_{i=1}^{m} \| r_i^k \| \right)
\]

\[
\leq \frac{1}{2} JC_{17} C_9 (\eta + \tau) \left( (\eta + \eta^2 + \tau^2) \left( \sum_{i=1}^{m} \| q_i^k \| \right)^{1/2} \right) \\
+ \frac{1}{2} JC_{17} C_9 (\eta + \tau) \left( (\eta + \eta^2 + \tau^2) \sum_{i=1}^{m} \| r_i^k \| \right)
\]

\[
+ JC_{17} C_9 (\eta + \tau) \left( (\eta + \eta^2 + \tau^2) \left( \sum_{i=1}^{m} \| r_i^k \| \right)^{1/2} \right)
\]

According to assumptions (A1) and (A2), there exists \( C_{18} > 0 \) such that: for any \( k \), \( \| p^k \| \leq C_{18} \).

\[
\left| \Delta u^{k+1} \cdot \xi^{j} \right| \leq C_{20} \left( \| u_{k, i}^{j} \| \| \Delta u_{i}^{k} \| + \eta \| p_{i}^{k} \| \right) \leq C_{20} (\eta + \tau) \| r_{i}^{k} \| \]

To prove (42), we set \( C_{20} = \max \left\{ \| \xi^{1} \|, \| \xi^{2} \|, \ldots, \| \xi^{J} \| \right\} \). Then we can get:

\[
\left| \Delta u^{k+1} \cdot \xi^{j} \right| \leq C_{20} \left( \| u_{k, i}^{j} \| \| \Delta u_{i}^{k} \| + \eta \| p_{i}^{k} \| \right) \leq C_{20} (\eta + \tau) \| r_{i}^{k} \| \]

According to assumptions (A1) and (A2), there exists \( C_{21} > 0 \) such that:

\[
\left| \frac{1}{2} \sum_{j=1}^{J} \sum_{p=1}^{P} \sum_{i=1}^{I} g_j'(w^{k+1/j})w_{p}^{j} g'(v_{p}^{j} B^{i+1/j} v_{p}^{j} g'(\tilde{\xi}_{i}^{j}) (\Delta u_{i}^{k+1} \cdot \xi^{j})^{2} \leq C_{21} \sum_{j=1}^{J} \sum_{p=1}^{P} \sum_{i=1}^{I} \left( \Delta u_{i}^{k+1} \cdot \xi^{j} \right)^{2}
\]

\[
\leq nJC_{20} (\eta + \tau) \sum_{i=1}^{m} \| r_{i}^{k} \| \leq C_{22} (\eta + \tau) \sum_{i=1}^{m} \| r_{i}^{k} \| ^{2}
\]
where \( C_{22} = nJC_{21}C_{20} \).

We continue to prove (43). According to assumptions (A1) and (A2), there exists \( C_{23} > 0 \) such that:

\[
\left| \sum_{j=1}^{n} \sum_{j=1}^{n} g_j(w^{k,j})v_j^k g_j^*(v_j^k) (U^{k+1,j} - U^{k,j}) \cdot \Delta v^k_p \right| \leq C_{23} \sum_{j=1}^{n} \sum_{j=1}^{n} \|U^{k+1,j} - U^{k,j}\| \|\Delta v^k_p\|
\]

Combining (15), (16) and (34), the above inequality can be rewritten as:

\[
\sum_{j=1}^{n} \sum_{j=1}^{n} g_j(w^{k,j})v_j^k g_j^*(v_j^k) (U^{k+1,j} - U^{k,j}) \cdot \Delta v^k_p \leq C_{23} C_0 \sum_{j=1}^{n} \sum_{j=1}^{n} (\eta + \tau) \|q_j^p\| \|\eta + \tau + \eta^2 + \tau^2\| \left( \sum_{i=1}^{m} \|r_i^k\|^2 \right)^{\frac{1}{2}}
\]

\[
\leq JC_{23} C_0 \sum_{j=1}^{n} (\eta + \tau) (\eta + \tau + \eta^2 + \tau^2) \|q_j^p\| \left( \sum_{i=1}^{m} \|r_i^k\|^2 \right)^{\frac{1}{2}}
\]

\[
\leq \frac{1}{2} JC_{23} C_0 (\eta + \tau) \left( \sum_{j=1}^{n} \|q_j^p\|^2 + \sum_{i=1}^{m} \|r_i^k\|^2 \right)
\]

\[
\leq C_{24} (\eta + \tau) \left( \sum_{j=1}^{n} \|q_j^p\|^2 + \sum_{i=1}^{m} \|r_i^k\|^2 \right)
\]

where \( C_{24} = \max \left\{ \frac{1}{2} JC_{23} C_0, \frac{n}{2} JC_{23} C_0 \right\} \).

Finally, we can choose \( C^* = \max \{C_{14}, C_{16}, C_{19}, C_{22}, C_{24} \} \) to complete the proof of the lemma.

Lemma 3.5 If (A1) and (A2) are valid, for the iteration process (15), there holds:

\[
E(w^{k+1}, V^{k+1}, U^{k+1}) \leq E(w^k, V^k, U^k) - \alpha \|p^k\|^2 - \beta \sum_{i=1}^{n} \|q_i^k\|^2 - \gamma \sum_{i=1}^{m} \|r_i^k\|^2
\]

where

\[
\alpha = \eta - \tau - C^* (\eta + \tau) \left( 2\eta + 2\tau + \eta^2 + \tau^2 + (\eta + \tau)^2 \right) - C^* (\eta + \tau)^2
\]

\[
\beta = \eta - \tau - C^* \left( \eta + \tau + \eta^2 + \tau^2 \right)^2 - C^* (\eta + \tau)^2 - C^* (\eta + \tau + (\eta + \tau)^2)
\]

\[
\gamma = \eta - \tau - C^* \left( \eta + \tau + \eta^2 + \tau^2 \right)^2 - 2C^* (\eta + \tau + \eta^2 + \tau^2)^2 - C^* (\eta + \tau)^2
\]

\[
- C^* (\eta + \tau) \left( \eta + \tau + \eta^2 + \tau^2 \right) - C^* (\eta + \tau) \left( \eta + \tau + \eta^2 + \tau^2 \right)^2
\]

(46)

and \( C^* \) is defined in Lemma 3.4.
Proof: Just notice that:
\[ |\tau_k p^k \cdot \Delta w^k| \leq \tau \left\| p^k \right\|^2, \]
\[ \sum_{i=1}^{m} |\gamma_{k,i,a} q_i^k \cdot \Delta u_i^k| \leq \sum_{i=1}^{m} |\gamma_{k,i,a} u_i^k| \left\| q_i^k \right\| \leq \tau \sum_{i=1}^{m} \left\| q_i^k \right\|^2, \]
\[ \sum_{i=1}^{m} |\gamma_{k,i,a} r_i^k \cdot \Delta u_i^k| \leq \sum_{i=1}^{m} \left\| \gamma_{k,i,a} u_i^k \right\| \left\| q_i^k \right\| \leq \tau \sum_{i=1}^{m} \left\| r_i^k \right\|^2. \]
Substitute all the conclusions of Lemma 3.4 into (33), we can complete the proof.

**Lemma 3.6** ([12]) Let \( f : \mathbb{R}^k \rightarrow \mathbb{R} \) be continuously differentiable. Assume that the number of the elements of the set \( \Omega = \{ x \mid f'_n(x) = 0 \} \) is finite. Then, if \( \{ x^k \} \) satisfies:
\[ \lim_{k \rightarrow \infty} \left\| x^k - x^k+1 \right\| = 0, \lim_{k \rightarrow \infty} \left\| f'_n(x^k) \right\| = 0 \]
Therefore,
\[ \lim_{k \rightarrow \infty} x^k = x^*, f_n(x^*) = 0 \quad (47) \]

**3.3. Proof of Theorems**
We first prove Theorem 2.1. It is easy to verify that if the formula below is satisfied,
\[ \eta < \frac{1 - \gamma}{C \left( \left( 2 + s + \gamma^2 \right)^4 + \left( 3 + s \right) \left( 2 + s + \gamma^2 \right)^2 + \left( 1 + s \right) \left( 5 + 3s + 2s^2 \right) \right)} \quad (48) \]
then \( \alpha, \beta \) and \( \gamma \) in (46) are all positive.
Consequently, according to (45), \( E(w^k, V^k, U^k) \) is monotonically decreasing.

Notice that the definition of \( E(w^k, V^k, U^k) \) shows that it is non-negative. So there exists \( E^* \geq 0 \) such that:
\[ \lim_{k \rightarrow \infty} E(w^k, V^k, U^k) = E^* \quad (49) \]
Combining (46) and (12)-(14) gives:
\[ \sum_{k=1}^{\infty} \left\| E_w \left( w^k, V^k, U^k \right) \right\|^2 = \sum_{k=1}^{\infty} \left\| p^k \right\|^2 < \infty, \]
\[ \sum_{k=1}^{\infty} \sum_{i=1}^{m} \left\| E_{v_i} \left( w^k, V^k, U^k \right) \right\|^2 = \sum_{k=1}^{\infty} \sum_{i=1}^{m} \left\| q_i^k \right\|^2 < \infty, \]
\[ \sum_{k=1}^{\infty} \sum_{i=1}^{m} \left\| E_{r_i} \left( w^k, V^k, U^k \right) \right\|^2 = \sum_{k=1}^{\infty} \sum_{i=1}^{m} \left\| r_i^k \right\|^2 < \infty. \]
Thus,
\[ \lim_{k \rightarrow \infty} \left\| E_w \left( w^k, V^k, U^k \right) \right\| = \lim_{k \rightarrow \infty} \left\| p^k \right\| = 0 \quad (50) \]
\[ \lim_{k \rightarrow \infty} \left\| E_{v_i} \left( w^k, V^k, U^k \right) \right\| = \lim_{k \rightarrow \infty} \left\| q_i^k \right\| = 0, \quad i = 1, 2, \ldots, n \quad (51) \]
\[ \lim_{k \rightarrow \infty} \left\| E_{r_i} \left( w^k, V^k, U^k \right) \right\| = \lim_{k \rightarrow \infty} \left\| r_i^k \right\| = 0, \quad i = 1, 2, \ldots, m \quad (52) \]
Combining (9)-(11), (50)-(52) and the formula used many times before, we have
\[ \| \Delta w^{k+1} \| \leq (\eta + \tau) \| \rho^k \| \]
\[ \| \Delta u_i^{k+1} \| \leq (\eta + \tau) \| q_i^k \|, \quad i = 1, 2, \ldots n \]
\[ \| \Delta u_i^{k+1} \| \leq (\eta + \tau) \| r_i^k \|, \quad i = 1, 2, \ldots m \]
Therefore,
\[ \lim_{k \to \infty} \| w^k - w^{k+1} \| = 0, \quad \lim_{k \to \infty} \| V^k - V^{k+1} \| = 0, \quad \lim_{k \to \infty} \| U^k - U^{k+1} \| = 0 \] (53)

To prove Theorem 2.2, simply applying Lemma 2.6 will obtain the result:
\[ \lim_{k \to \infty} w^k = w^*, \quad \lim_{k \to \infty} V^k = V^*, \quad \lim_{k \to \infty} U^k = U^* \] (54)
\[ \lim_{k \to \infty} E_w(w^k, V^k, U^k) = 0, \]
\[ \lim_{k \to \infty} E_{v_i}(w^k, V^k, U^k) = 0, i = 1, 2, \ldots n \] (55)
\[ \lim_{k \to \infty} E_{u_i}(w^k, V^k, U^k) = 0, i = 1, 2, \ldots m \]

According to Lemma 3.5, we know that \((w^*, V^*, U^*)\) is a local minimum. So far, we have completed the proof of theorems in Section 2.

4. Conclusion
This paper mainly gives a theoretical proof of the convergence of the momentum BP algorithm with double hidden layers. We prove that the momentum BP algorithm with double hidden layers is convergent with carefully chosen momentum coefficients. In future research, we will try to prove that the momentum BP algorithm converges with more hidden layers.

References
[1] T. L. Fine, S. Mukherje, Parameter convergence and learning curves for neural networks, J. Neural Computation, 1999, 11: 747-769.
[2] W. Finnoff, Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima, Neural Comput., 6 (1994), 285-295.
[3] X. Yu, N. K. Loh, W. C. Miller, A new acceleration technique for the back-propagation algorithm, in: IEEE International Conference on Neural Networks, vol. 3, 1993, pp. 1157–1161.
[4] E. Istook, T. Martinez, Improved backpropagation learning in neural networks with windowed momentum, International Journal of Neural System 12 (2002) 303–318.
[5] C. Yu, B. Liu, A backpropagation algorithm with adaptive learning rate and momentum coefficient, Proceedingsofthe2002InternationalJointConferenceonNeuralNetworks, vol. 2, 2002, pp. 1218–1223.
[6] H. M. Shao, G. F. Zheng, A new BP algorithm with adaptive momentum for FNNs training, in: 2009 WRI Global Congress on Intelligent Systems, vol. 4, 2009, pp. 16–20.
[7] N. M. Zhang, W. Wu, G.F. Zheng, Convergence of gradient method with momentum for two-layer feedforward neural networks, IEEE T. Neural Netw., 17:2 (2006), 522-525.
[8] N. M. Zhang, An online gradient method with momentum for two-layer feedforward neural networks. Applied Mathematics and Computation. 2009, 212(2): 488-498.
[9] W. Wu, N. M. Zhang, Z. X. Li et al. Convergence of Gradient Method with Momentum for Back-
Propagation Neural Networks. Journal of Computational Mathematics, 2008, 26: 613–623.

[10] A. Bhaya, E. Kaszkurewicz, Steepest descent with momentum for quadratic functions is a version of the conjugate gradient method, Neural Networks 17 (2004) 65–71.

[11] M. Torii and M. T. Hagan, Stability of steepest descent with momentum for quadratic functions, IEEE T. Neural Networ., 13:3 (2002), 752-756.

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