Bulletin of National University of Uzbekistan: Mathematics and Natural Sciences

Volume 2 | Issue 2 Article 1

6-6-2019

Convergence analysis of feedforward neural networks with backpropagation

Avazjon Marakhimov
National University of Uzbekistan, avaz.marakhimov@yandex.ru

Kabul Khudaybergenov
National University of Uzbekistan, kabul85@mail.ru

Recommended Citation

Marakhimov, Avazjon and Khudaybergenov, Kabul (2019) "Convergence analysis of feedforward neural networks with backpropagation," *Bulletin of National University of Uzbekistan: Mathematics and Natural Sciences*: Vol. 2: Iss. 2, Article 1.
DOI: https://doi.org/10.56017/2181-1318.1022

CONVERGENCE ANALYSIS OF FEEDFORWARD NEURAL NETWORKS WITH BACKPROPAGATION

Marakhimov A. R., Khudaybergenov K. K.
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: avaz.marakhimov@yandex.ru, kabul85@mail.ru

Abstract

In this paper, convergence analysis properties of online gradient training for backpropagation algorithm for feedforward neural networks with a two hidden layer is studied. We assume that in every training cycle, every training pattern in the training dataset is fed in a stochastic form to the feedforward multilayer neural network exactly once. In this study, we give a weak and strong convergence properties for the training approaches, indicating that the gradient of the error function goes to zero and the weights goes to a fixed point value, respectively. First, we give convergence result for completely stochastic order approach and then follows for special stochastic order approach. The conditions on the activation function of the network and the training rate to guarantee the convergence are relaxed compared with the existing results. Convergence properties in the current paper are studied for sigmoidal activation function type, however this results are also valid for other type of activation functions.

Keywords: neural networks, backpropagation method, online gradient, weak convergence, strong convergence.

Mathematics Subject Classification (2010): 97R40, 92B20.

1. Introduction

The past two decades have seen an enormous change in the field of artificial neural networks and its application. Especially, there has been a considerable growth of progress in feed forward neural network with a number of layers [1]-[7]. The application areas of artificial neural network include wide topics in computational intelligence, biology, decision science, medicine, finance, intelligent information processing. Backpropagation method is one of the widely used training method for feedforward artificial neural networks. This training method was introduced by Werbos [6] in 1974. At the same time, we can see there are other authors who proposed this method simultaneously [8]-[9]. Algorithm of this method can be implemented in two separate approaches: batch updating and online updating. Implementing the batch updating parameter approach with a standard gradient method, the weight correction operation is performed over all the training samples before updating the current weight. However, the online updating variant performs the updates directly after every training sample is passed.

There exists three cases for online training of backpropagation with a online gradient method (OGM): OGM-CS (completely stochastic order), OGM-SS (special stochastic order), OGM-F (fixed order). For OGM-CS case, at every training step,

a one pattern is chosen randomly from the training dataset and fed to the neural network [10]. For OGM-SS approach, on every training cycle, every pattern in the training dataset is fed in a stochastic order to the neural network exactly once [11]. In OGM-F case, in every training cycle, every pattern in the training dataset is fed in a fixed order to the neural network exactly once [12].

In fact, we can find in some papers that convergence properties for OGM-CS approach are asymptotic convergence results with a probabilistic character as the volume of training dataset tends to infinity [13]. Also, deterministic convergence properties for OGM-SS and OGM-F approaches are deeply studied in [14]. We can see that the training method OGM-SS approach with stochastic character makes use of deterministic convergence character. To obtain proves of the convergence properties for OGM-F approach is much more simply compared with OGM-SS approach.

In general, to guarantee the convergence for OGM-CS approach, it is supposed that the training rate η_m for updating network parameters must satisfy the assumptions $\sum_{m=1}^{\infty}\eta_m = \infty$ and $\sum_{m=1}^{\infty}\eta_m^2 < \infty$. Furthermore, in [13] was given an additional assumption $\lim_{m\to\infty}\eta_m/\eta_{m+1} = 1$ for OGM-F approach.

In order to show the strong convergence, which implies the weights values converges to a fixed point, in [15] was introduced an additional assumption that the number of the stationary points of the error function is finite. In [16] a more relaxed condition has been used which is the gradient of the error function has at most countably infinite number of stationary points.

The objective of this paper is to give the weak and strong convergence properties both for OGM-F and OGM-SS approaches. This implies that the gradient of the error function infinitely approaches to 0 and the weight parameters obtains a fixed point value, respectively.

In the current note, we extend the results which is obtained in [17]. We implement a four layered feedforward neural network, which has a two hidden layer.

The paper is organised in following form. In Section 1, description of the problem statement is given. In Section 2, types of online updating methods are introduced: OGM-F, OGM-SS. Theorem about the convergence are presented in Section 3 and its proofs are given in Section 4. And some conclusions are given at the end of the paper.

1 Problem statement

Let us consider a feedforward neural network with four layer architecture (2-hidden layers). The number of neurons in the input, 1-hidden, 2-hidden and output layers in network is equal to s, p, n, and 1, respectively. We suppose the training dataset is $\{x^j, O^j\}_{j=1}^{J} \subset \mathbb{R}^s \times \mathbb{R}$ where x^j and O^j are the input and the corresponding desired output of the $j-$th pattern (sample), respectively. Let $V = (v_{i,j})_{p\times s}$ be the weight matrix connecting the input and the 1-hidden layer, and write $v_i = (v_{i1}, v_{i2}, \ldots, v_{is})^T$ for $i = 1, 2, \ldots, p$ and $B = (b_{i,j})_{n\times p}$ - be the weight matrix connecting the 1-hidden and the 2-hidden layer, and write $b_i = (b_{i1}, b_{i2}, \ldots, b_{ip})^T$ for

$i = 1, 2, \ldots, n$. The weight vector connecting the 2-hidden and output layer is denoted as $d = (d_1, d_2, \ldots, d_n)^T \in \mathbb{R}^n$. We simplify the presentation by combining the weight matrixes V, B and the weight vector d, and write in one vector as follows $w = \left(d^T, b_1^T, b_2^T, \ldots, b_n^T, v_1^T, v_2^T, \ldots, v_p^T\right)^T \in \mathbb{R}^{p(s+n)+n}$. Let $h, g, f : \mathbb{R} \to \mathbb{R}$ be given activation functions for the 1-hidden, 2-hidden and output layers, respectively. For convenience, we write the following vector valued function as the following forms

$$H(z) = (h(z_1), h(z_2), \ldots, h(z_n))^T, \ \ \forall z \in \mathbb{R}^n, \tag{1}$$

$$G(z) = (g(z_1), g(z_2), \ldots, g(z_n))^T, \ \ \forall z \in \mathbb{R}^n. \tag{2}$$

For a given input $x \in \mathbb{R}$, the final output from the neural network can be calculated as follows

$$y = f(d \cdot H(B \cdot G(Vx))). \tag{3}$$

For any fixed weight parameter w, the error of the neural network is defined as following form

$$E(w) = \frac{1}{2}\sum_{j=1}^{J}\left(O^j - f\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right)\right)^2 = \sum_{j=1}^{J} f_j\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right), \tag{4}$$

where $f_j(t) = \frac{1}{2}(O^j - f(t))^2$, $j = 1, 2, \ldots, J$, $t \in \mathbb{R}$. The gradient formulas of the error function with regarding to parameters d, b_i and v_i are, respectively, given by

$$E_d(w) = -\sum_{j=1}^{J}\left(O^j - y^j\right) f'\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right) H\left(B \cdot G\left(Vx^j\right)\right)$$

$$= \sum_{j=1}^{J} {f'}_j\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right) H\left(B \cdot G\left(Vx^j\right)\right), \tag{5}$$

$$E_{b_i}(w) = -\sum_{j=1}^{J}\left(O^j - y^j\right) f'\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right) d_i h'\left(b_i \cdot g\left(v_i \cdot x^j\right)\right) G\left(Vx^j\right)$$

$$= \sum_{j=1}^{J} {f'}_j\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right) d_i h'\left(b_i \cdot g\left(v_i \cdot x^j\right)\right) G\left(Vx^j\right), \tag{6}$$

$$E_{v_i}(w) = -\sum_{j=1}^{J}\left(O^j - y^j\right) f'\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right) d_i h'\left(b_i \cdot g\left(v_i \cdot x^j\right)\right) b_i g'\left(v_i \cdot x^j\right) x^j$$

$$= \sum_{j=1}^{J} {f'}_j\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right) d_i h'\left(b_i \cdot g\left(v_i \cdot x^j\right)\right) b_i g'\left(v_i \cdot x^j\right) x^j, \tag{7}$$

where

$$y^j = f\left(d \cdot H\left(B \cdot G\left(Vx^j\right)\right)\right), \ \ i = 1, 2, \ldots, n; \ \ j = 1, 2, \ldots, J. \tag{8}$$

Write

$$E_B(w) = \left(E_{b_1}(w)^T, E_{b_2}(w)^T, \ldots, E_{b_s}(w)^T\right)^T, \tag{9}$$

$$E_V(w) = \left(E_{v_1}(w)^T, E_{v_2}(w)^T, \ldots, E_{v_p}(w)^T\right)^T, \tag{10}$$

$$E_w(w) = \left(E_d(w)^T, E_B(w)^T, E_V(w)^T\right)^T. \tag{11}$$

2 Approaches to online training of neural networks

As we mention above, there two approaches to train the network. Let us consider as first variant that the training patterns are fed to the network in a fixed form (OGM-F) in the training procedure. Consequently, beginning from some initial value w^0, we may continue to update it iteratively by the formulas which is given below

$$d^{mJ+j+1} = d^{mJ+j} + \Delta_j d^{mJ+j}, \tag{12}$$

$$b_i^{mJ+j+1} = b_i^{mJ+j} + \Delta_j b^{mJ+j}, \tag{13}$$

$$v_i^{mJ+j+1} = v_i^{mJ+j} + \Delta_j v_i^{mJ+j}, \tag{14}$$

where

$$\begin{aligned}\Delta_k d^{mJ+j} &= \eta_m \left(O^k - y^{mJ+j,k}\right) f'\left(d^{mJ+j} \cdot H^{mJ+j,k}\right) H^{mJ+j,k} \\ &= -\eta_m {f'}_k \left(d^{mJ+j} \cdot H^{mJ+j,k}\right) H^{mJ+j,k},\end{aligned} \tag{15}$$

$$\begin{aligned}\Delta_k b_i^{mJ+j} &= \eta_m \left(O^k - y^{mJ+j,k}\right) f'\left(d^{mJ+j} \cdot H^{mJ+j,k}\right) d_i^{mJ+j} \\ &\times h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^k\right)\right) G^{mJ+j,k} \\ = -\eta_m {f'}_k &\left(d^{mJ+j} \cdot H^{mJ+j,k}\right) d_i^{mJ+j} h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^k\right)\right) \cdot G^{mJ+j,k}\end{aligned} \tag{16}$$

$$\begin{aligned}\Delta_k v_i^{mJ+j} &= \eta_m \left(O^k - y^{mJ+j,k}\right) f'\left(d^{mJ+j} \cdot H^{mJ+j,k}\right) d_i^{mJ+j} \\ &\times h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^k\right)\right) b_i^{mJ+j} g'\left(v_i^{mJ+j} \cdot x^k\right) x^k \\ = -\eta_m {f'}_k &\left(d^{mJ+j} \cdot H^{mJ+j,k}\right) d_i^{mJ+j} h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^k\right)\right) \\ &\times b_i^{mJ+j} g'\left(v_i^{mJ+j} \cdot x^k\right) x^k,\end{aligned} \tag{17}$$

$$\begin{gathered}G^{mJ+j,k} = G\left(V^{mJ+j} x^k\right), \quad H^{mJ+j,k} = H\left(B^{mJ+j} \cdot G^{mJ+j,k}\right), \\ y^{mJ+j,k} = f\left(d^{mJ+j} \cdot H^{mJ+j,k}\right), \\ m \in \mathbb{N}; \quad i = 1, 2, \ldots, n; \quad j, k = 1, 2, \ldots, J.\end{gathered} \tag{18}$$

here, η_m - the parameter for the learning rate. This value may be varied after every cycle of the training process.

As for a special stochastic order (OGM-SS), the training patterns can be organized as follows: For the mth training cycle, let $\left\{x^{m,1}, x^{m,2}, \ldots, x^{m,J}\right\}$ be a stochastic permutation of the given dataset $\left\{x^1, x^2, \ldots, x^J\right\}$ Similarly, the weight parameters (12), (13) and (14) are iteratively updated in the following form

$$d^{mJ+j+1} = d^{mJ+j} + \Delta_j^m d^{mJ+j}, \tag{19}$$

$$b_i^{mJ+j+1} = b_i^{mJ+j} + \Delta_j^m b_i^{mJ+j}, \tag{20}$$

$$v_i^{mJ+j+1} = v_i^{mJ+j} + \Delta_j^m v_i^{mJ+j}, \tag{21}$$

where

$$\begin{aligned}\Delta_k^m d^{mJ+j} &= \eta_m \left(O^k - y^{mJ+j,m,k}\right) f'\left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right) H^{mJ+j,m,k} \\ &= -\eta_m {f'}_k \left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right) H^{mJ+j,m,k},\end{aligned} \tag{22}$$

$$\begin{aligned}\Delta_k^m b_i^{mJ+j} &= \eta_m \left(O^k - y^{mJ+j,m,k}\right) f'\left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right) \\ &\times d_i^{mJ+j} h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,m,k} \cdot x^{m,k}\right)\right) G^{mJ+j,m,k} \\ &= -\eta_m {f'}_k \left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right) d_i^{mJ+j} h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,m,k} \cdot x^{m,k}\right)\right) G^{mJ+j,m,k},\end{aligned} \tag{23}$$

$$\begin{aligned}\Delta_m^k v_i^{mJ+j} &= \eta_m \left(O^k - y^{mJ+j,m,k}\right) f'\left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right) \\ &\times d_i^{mJ+j} h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^{m,k}\right)\right) b_i^{mJ+j} g'\left(v_i^{mJ+j,k} \cdot x^{m,k}\right) x^{m,k} \\ &= -\eta_m {f'}_k \left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right) d_i^{mJ+j} h'\left(b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^{m,k}\right)\right) \\ &\times b_i^{mJ+j} g'\left(v_i^{mJ+j,k} \cdot x^{m,k}\right) x^{m,k},\end{aligned} \tag{24}$$

$$G^{mJ+j,m,k} = G\left(V^{mJ+j} x^{m,k}\right), \quad H^{mJ+j,m,k} = H\left(B^{mJ+j} \cdot G^{mJ+j,m,k}\right),$$

$$y^{mJ+j,m,k} = f\left(d^{mJ+j} \cdot H^{mJ+j,m,k}\right), \tag{25}$$

$$m \in \mathbb{N}; \;\; i = 1, 2, \ldots, n; \;\; j, k = 1, 2, \ldots, J. \tag{26}$$

3 Main results

For $\forall x \in \mathbb{R}^n$ consider Euclidean norm on $\mathbb{R}^n$

$$\|x\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}, \;\; x = (x_1, \ldots, x_n).$$

Let the set $\Omega_0 = \{w \in \Omega : \;\; E_w(w) = 0\}$ be the stationary point set of the error function $E(w)$, where $\Omega \in \mathbb{R}^{p(s+n)+n}$ is a bounded region satisfying the assumption

(A3) which is given below. Let $\Omega_{0,s} \in \mathbb{R}$ be the projection of Ω_0 onto the $s-$th coordinate axis, that is,

$$\Omega_{0,s} = \left\{ w_s \in \mathbb{R} : w = \left(w_1, w_2, \ldots, w_{p(s+n)+n}\right)^T \in \Omega_0 \right\}, \tag{27}$$

where $s = 1, 2, \ldots, p\,(s+n)+n$. To obtain the convergence results of the algorithm, there are necessity for the following assumptions which given below.

(A1) $h'\,(t)$, $g'\,(t)$ and $f'\,(t)$ are Lipschitz continuous on any bounded closed interval;

(A2) The training parameter, $\eta_m > 0$, $\sum_{m=0}^{\infty} \eta_m = \infty$, $\sum_{m=0}^{\infty} \eta_m^2 = \infty$;

(A3) There exists a bounded open set $\Omega \in \mathbb{R}^n$ such that $\{w^m\} \subset \Omega$ $(m \in N)$;

(A3') There exists a bounded open set $\Omega' \in \mathbb{R}^n$ such that $\{d^m\} \subset \Omega'$ $(m \in N)$ and the derivative of the activation function h, g in (1), (2) are uniformly bounded and Lipschitz continuous on $\mathbb{R}$, respectively;

(A4) $\Omega_{0,s}$ does not contain any interior point for every $s = 1, 2, \ldots, p\,(s+n)+n$.

Theorem 1. *Assume that conditions (A1)-(A3) are valid. Then, starting from an arbitrary initial value w^0, the learning sequence w^m defined by* (12), (13), (14) *or by* (19), (20) *and* (21) *satisfies the following weak convergence*

$$\lim_{m\to\infty} \|E_w\,(w^m)\| = 0. \tag{28}$$

Moreover, if assumptions (A1)-(A4) are valid, there holds the strong convergence: There exists $w^* \in \Omega_0$ such that

$$\lim_{m\to\infty} w^m = w^*. \tag{29}$$

Let us make three the following remarks on the convergence property: (1) We claim that the weak convergence remains valid if the activation functions h and g of the hidden layers are a commonly used sigmoid function and assumptions (A3') (instead of (A3)) and (A2) are valid. This is due to the fact that the sigmoid functions h and g are uniformly bounded on $\mathbb{R}$ and that (46) is valid even if the weight vectors $v_i = 1, 2, \ldots, s$ are unbounded. (2) In the numerical analysis of an iterative method for a class of nonlinear problems, the iterative sequence is often required to be bounded in order to prove its convergence. This is what we do in conditions (A3) and (A3'). We mention that the weights will be automatically bounded in the network training with the help of a penalty term [12]. (3) For the strong convergence, our condition (A4) on Ω_0 allows it to be finite set, countably infinite set, nowhere dense set or even some uncountable dense set. Hence, the corresponding assumptions that the set Ω_0 contains finite points and at most countably infinite points in [15] and [16], respectively, are special cases of assumption (A4). This relaxed condition makes it much easier to verify the strong convergence in practice.

4 Proofs

We show in more detail the convergence proof for OGM-F approach in the Sections 4.1-4.2. Then, in Section 4.3, we briefly point out how to make extention to OGM-SS approach.

4.1 Convergence analysis for OGM-F approach

Firstly, we present the following necessary lemmas for the convergence analysis.

Lemma 1. *[17] Let $q(x)$ be a function defined on a bounded closed interval $[a, b]$ such that $q'(x)$ is Lipschitz continuous with Lipschitz constant $K > 0$. Then, $q'(x)$ is differentiable almost everywhere in $[a, b]$ and*

$$|q''(x)| \leq K. \tag{30}$$

Moreover, there exists a constant $C > 0$ such that

$$q(x) \leq q(x_0) + q'(x_0)(x - x_0) + C(x - x_0)^2, \quad \forall x_0, x \in [a, b]. \tag{31}$$

Lemma 2. *[17] Suppose that the learning rate η_m satisfies (A2) and that the sequence $\{a_m\}$ $(m \in \mathbb{N})$ satisfies $a_m \geq 0$, $\sum_{m=0}^{\infty} \eta_m a_m^{\beta} < \infty$ and $|a_{m+1} - a_m| \leq \mu\eta_m$ for some positive constants β and μ. Then we have*

$$\lim_{m\to\infty} a_m = 0. \tag{32}$$

Lemma 3. *Let $\{b_m\}$ be a bounded sequence satisfying $\lim_{m\to\infty}(b_{m+1} - b_m) = 0$. Write $\gamma_1 = \lim_{n\to\infty}\inf_{m>n} b_m$, $\gamma_2 = \lim_{n\to\infty} sup_{m>n} b_m$ and $S = \{a \in \mathbb{R} :$ There exists a subsequence $\{b_{i_k}\}$ of $\{b_m\}$ such that $b_{i_k} \to a$ as $k \to \infty\}$. Then we have*

$$S = [\gamma_1, \gamma_2]. \tag{33}$$

Proof. It is obvious that $\gamma_1 \leq \gamma_2$ and $S = [\gamma_1, \gamma_2]$. If $\gamma_1 = \gamma_2$, then (33) follows simply from $\lim_{m\to\infty} b_m = \gamma_1 = \gamma_2$. Let us consider the case $\gamma_1 < \gamma_2$ and proceed to prove that $S \supseteq [\gamma_1, \gamma_2]$.

For any $\forall a \in (\gamma_1, \gamma_2)$, there exists $\varepsilon > 0$ such that $(a - \varepsilon, a + \varepsilon) \subseteq (\gamma_1, \gamma_2)$. Noting that $\lim_{m\to\infty}(b_{m+1} - b_m) = 0$, we observe that b_m travels to and from between γ_1 and γ_2 with very small pace for all large enough m. Hence, there must be infinite number of points of the sequence b_m falling into $(a - \varepsilon, a + \varepsilon)$. This implies $a \in S$ and thus $(\gamma_1, \gamma_2) \subseteq S$. Furthermore, $(\gamma_1, \gamma_2) \subseteq S$ immediately leads to $[\gamma_1, \gamma_2] \subseteq S$. This completes the proof.

Let the weight sequence $\left\{w^{mJ+j}\right\}$ $(m \in \mathbb{N},\ j = 1, 2, \ldots, J)$ be generated by (12), (13) and (14). We define the notations below as follows:

$$R^{m,j} = \Delta_j d^{mJ+j} - \Delta_j d^{mJ}, \tag{34}$$

$$Q_i^{m,j} = \Delta_j b_i^{mJ+j} - \Delta_j b_i^{mJ}, \tag{35}$$

$$r_i^{m,j} = \Delta_j v_i^{mJ+j} - \Delta_j v_i^{mJ}, \tag{36}$$

$$\xi^{m,l} = d^{mJ+l} - d^{mJ} = \sum_{j=1}^{l} \Delta_j d^{mJ+j} = \sum_{j=1}^{l} \Delta_j d^{mJ} + \sum_{j=1}^{l} R^{m,j}, \tag{37}$$

$$\tau_i^{m,l} = b_i^{mJ+l} - b_i^{mJ} = \sum_{j=1}^{l} \Delta_j b_i^{mJ+j} = \sum_{j=1}^{l} \Delta_j b_i^{mJ} + \sum_{j=1}^{l} Q^{m,j}, \tag{38}$$

$$\phi_i^{m,l} = v_i^{mJ+l} - v_i^{mJ} = \sum_{j=1}^{l} \Delta_j v_i^{mJ+j} = \sum_{j=1}^{l} \Delta_j v_i^{mJ} + \sum_{j=1}^{l} r_i^{mJ}, \tag{39}$$

$$\psi^{m,l,j} = G^{mJ+l,j} - G^{mJ,j}, \tag{40}$$

$$\varphi^{m,l,j} = H^{mJ+l,j} - H^{mJ,j}, \tag{41}$$

$m \in \mathbb{N};\ \ j = 1, 2, \ldots, J;\ \ l = 1, 2, \ldots, J,\ \ i = 1, 2, \ldots, n.$

Then, (12), (13) and (14) can be rewritten as follows

$$d^{mJ+j} = d^{mJ} + \sum_{k=1}^{j} \left(\Delta_k d^{mJ} + R^{m,k}\right), \tag{42}$$

$$b_i^{mJ+j} = b_i^{mJ} + \sum_{k=1}^{j} \left(\Delta_k b_i^{mJ} + Q_i^{m,k}\right), \tag{43}$$

$$v_i^{mJ+j} = v_i^{mJ} + \sum_{k=1}^{j} \left(\Delta_k v_i^{mJ} + r_i^{m,j}\right). \tag{44}$$

Let constants C_1 and C_2 be defined by (cf. assumption (A3))

$$\max_{1\le j\le J} \left\{\left\|x^j\right\|, \left|O^j\right|\right\} = C_1,\ \ \sup_{m\in\mathbb{N}} \|w^m\| = C_2. \tag{45}$$

By assumption (A1), $f'_j(t)$ also satisfies the Lipschitz condition for $j = 1, 2, \ldots, J$. Furthermore, $g(t)$, $f(t)$, $h(t)$ and $f_j(t)$ are all uniformly continuous on any bounded closed interval. □

Lemma 4. *Let conditions (A1) and (A3) be valid, and let the sequence* $\left\{w^{mJ+j}\right\}$ *be generated by* (12), (13) *and* (14). *Then there are constants* $C_3 - C_{10}$ *such that*

$$\left\|G^{mJ+j,k}\right\| \le C_3,\ \ \left\|H^{mJ+j,k}\right\| \le C_4, \tag{46}$$

$$\left\|\xi^{m,l}\right\| \le C_5\eta_m,\ \ \left\|\psi^{m,l,j}\right\| \le C_6\eta_m,\ \ \left\|\varphi^{m,l,j}\right\| \le C_7\eta_m, \tag{47}$$

$$\left\|R^{m,l}\right\| \le C_8\eta_m^2,\ \ \left\|Q_i^{m,j}\right\| \le C_9\eta_m^2,\ \ \left\|r_i^{m,j}\right\| \le C_{10}\eta_m^2, \tag{48}$$

where $m \in \mathbb{N};\ \ j, k = 1, 2, \ldots, J;\ \ l = 1, 2, \ldots, J,\ \ i = 1, 2, \ldots, n.$

Proof. According to (45), we have

$$\left|v_i^{mJ+j} \cdot x^k\right| \leq \left\|v_i^{mJ+j}\right\| \left\|x^k\right\| \leq C_1 C_2 \equiv D_1. \tag{49}$$

Thus, there exists a positive constant $C_{3,1}$ such that

$$\max_{|t| \leq D_1} |g(t)| = C_{3,1}, \tag{50}$$

$$\left\|G^{mJ+j,k}\right\| = \left\|G\left(V^{mJ+j} x^k\right)\right\| \leq \sqrt{n} C_{3,1} \equiv C_3. \tag{51}$$

Also, according to (45) and (50), we have

$$\left|b_i^{mJ+j} \cdot g\left(v_i^{mJ+j,k} \cdot x^k\right)\right| \leq \left\|b_i^{mJ+j}\right\| \cdot C_{3,1} \leq C_2 C_{3,1} \equiv D_2. \tag{52}$$

Thus, there exists a positive constant $C_{4,1}$ such that

$$\max_{|t| \leq D_2} |h(t)| = C_{4,1}, \tag{53}$$

$$\left\|H^{mJ+j,k}\right\| = \left\|H\left(B^{mJ+j} \cdot G\left(V^{mJ+j} x^k\right)\right)\right\| \leq \sqrt{p} C_{4,1} \equiv C_4. \tag{54}$$

It follows from (45) and (54) that

$$\left|d^{mJ+j} \cdot H^{mJ+j,k}\right| \leq \left\|d^{mJ+j}\right\| \left\|H^{mJ+j,k}\right\| \leq C_2 C_4 \equiv D_3. \tag{55}$$

Then there is a positive constant $C_{5,1}$ such that

$$\max_{|t| \leq D_3} \left|f'_j(t)\right| = C_{5,1}. \tag{56}$$

In addition, the combination of (A1), (15), (46) and (49) gives

$$\left\|\xi^{m,l}\right\| = \left\|d^{mJ+l} - d^{mJ}\right\| = \left\|\sum_{j=1}^{l} \Delta_j d^{mJ+j}\right\| \leq C_5 \eta_m, \tag{57}$$

where $C_5 = J C_{5,1} C_4$.

Employing (49), we find that

$$\max_{|t| \leq D_1} |g'(t)| = C_{6,1}, \tag{58}$$

where $C_{6,1}$ is a positive constant. Moreover, we observe that

$$\left\|\psi^{m,l,j}\right\| = \left\|G^{mJ+l,j} - G^{mJ,j}\right\| \leq \max_{1 \leq j \leq n} |g'(t_i)| \left\|x^j\right\| \sum_{i=1}^{n} \left\|\phi_i^{m,l}\right\|$$

$$\leq \max_{1 \leq j \leq n} |g'(t_i)| \left\|x^j\right\| \sum_{i=1}^{n} \sum_{k=1}^{l} \left\|\Delta_k v_i^{mJ+k}\right\| \leq C_6 \eta_m, \tag{59}$$

where $C_6 = nlC_{5,1}C_{6,1}\max_{1\le i\le n}|g'(t_i)|\,\|x^j\|\sup_{m\in\mathbb{N}}\|w^n\|\max_{1\le k\le j}\|x^k\|$, in which $t_i = v_i^{mJ}\cdot x^j+\theta_i\left(v_i^{mJ+l}-v_i^{mJ}\right)\cdot x^j$, $\theta_i\in(0,\,1)$, and $|t_i|\le\left|v_i^{mJ}\cdot x^j\right|+\left|\theta_i\left(v_i^{mJ+l}-v_i^{mJ}\right)\cdot x^j\right|\le 3C_1C_2$. By virtue of (A1), we see that $|g'(t_i)|$ $(i=1,\,2,\ldots,\,n)$ is bounded.

In the same way, one can show the existence of a constant $C_7>0$ such that

$$\left\|\varphi^{m,l,j}\right\|\le C_7\eta_m^2. \tag{60}$$

Combining $f'_j(t)$'s Lipschitz continuity, (45) and (46), we have

$$\begin{aligned}\left|f'_j\left(d^{mJ+j}\cdot H^{mJ+j,j}\right)-f'_j\left(d^{mJ}\cdot H^{mJ+j,j}\right)\right|&\le L\left|d^{mJ+j}\cdot H^{mJ+j,j}-d^{mJ}\cdot H^{mJ+j,j}\right|\\ &\le L\left\|\xi^{m,l}\right\|\left\|H^{mJ+j,j}\right\|\le LC_4\left\|\xi^{m,l}\right\|,\end{aligned} \tag{61}$$

$$\begin{aligned}\left|f'_j\left(d^{mJ}\cdot H^{mJ+j,j}\right)-f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\right|&\le L\left|d^{mJ}\cdot H^{mJ+j,j}-d^{mJ}\cdot H^{mJ,j}\right|\\ &\le L\left\|d^{m,j}\right\|\left\|\varphi^{m,j,j}\right\|\le LC_2\left\|\varphi^{m,j,j}\right\|,\end{aligned} \tag{62}$$

where $L>0$ is the Lipschitz constant. By the definition of $R^{m,j}$, we see that

$$\begin{aligned}R^{m,j}&=\Delta_j d^{mJ+j}-\Delta_j d^{mJ}\\ &=-\eta_m\left(f'_j\left(d^{mJ+j}\cdot H^{mJ+j,j}\right)H^{mJ+j,j}-f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)H^{mJ,j}\right)\\ &=-\eta_m\left[f'_j\left(d^{mJ+j}\cdot H^{mJ+j,j}\right)\varphi^{m,j,j}\right.\\ &\quad+\left(f'_j\left(d^{mJ+j}\cdot H^{mJ+j,j}\right)-f'_j\left(d^{mJ}\cdot H^{mJ+j,j}\right)\right)H^{mJ,j}\\ &\quad\left.+\left(f'_j\left(d^{mJ}\cdot H^{mJ+j,j}\right)-f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\right)H^{mJ,j}\right].\end{aligned} \tag{63}$$

Therefore, it follows from (46), (47), (61), and (62) that

$$\left\|R^{m,j}\right\|\le\eta_m\left(LC_4^2\left\|\xi^{m,j}\right\|+\left(C_{5,1}+LC_2C_4\right)\left\|\varphi^{m,j,j}\right\|\right)\le C_8\eta_m^2, \tag{64}$$

where $C_8=\max\left\{LC_4^2C_5,\ \left(C_{5,1}+LD_3\right)C_7\right\}$.

Similarly, we can show the existence of a constant $C_9, C_{10}>0$ such that

$$\left\|Q_i^{m,j}\right\|\le C_9\eta_m^2,\quad\left\|r_i^{m,j}\right\|\le C_{10}\eta_m^2. \tag{65}$$

□

The next lemma reveals an almost monotonicity of the error function during the training process.

Lemma 5. *Let the sequence $\left\{w^{mJ+j}\right\}$ be generated by* (12), (13) *and* (14). *Under assumptions (A1) and (A3), there holds*

$$E\left(w^{(m+1)J}\right)\le E\left(w^{mJ}\right)-\eta_m\left\|E_w\left(w^{mJ}\right)\right\|^2+C_{11}\eta_m^2,\ (m=0,\,1,\ldots) \tag{66}$$

where $C_{11}>0$ is a constant independent of m and η_m.

Proof. By virtue of assumption (A1) and Lemma 1, we know that the derivative $h''\left(b_i^{mJ}\cdot g\left(v_i^{mJ}x^j\right)+t\left(\tau_i^{mJ}\cdot x^j\right)\right)$ is integrable almost everywhere on [0, 1] and

$$f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)d^{mJ}\cdot\varphi^{m,J,j}=f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\sum_{i=1}^{n}d_i^{mJ}h'\left(b_i^{mJ}\cdot g\left(v_i^{mJ}\cdot x^j\right)\right)\tau_i^{m,J}\cdot x^j$$

$$+f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\sum_{i=1}^{n}d_i^{mJ}\left(\tau_i^{m,l}\cdot x^j\right)^2$$

$$\times\int_0^1(1-t)h''\left(b_i^{mJ}\cdot g\left(v_i^{mJ}x^j\right)+t\left(\tau_i^{mJ}\cdot x^j\right)\right)dt. \tag{67}$$

By virtue of Lemma 1, (16), (17) and (67), there is a constant $C_{12}>0$ such that

$$f_j\left(d^{(m+1)J}\cdot H^{(m+1)J,j}\right)\le f_j\left(d^{mJ}\cdot H^{mJ,j}\right)+f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)$$

$$\times\left(d^{(m+1)J}\cdot H^{(m+1)J,j}-d^{mJ}\cdot H^{mJ,j}\right)+C_9\left(d^{(m+1)J}\cdot H^{(m+1)J,j}-d^{mJ}\cdot H^{mJ,j}\right)^2$$

$$=f_j\left(d^{mJ}\cdot H^{mJ,j}\right)$$

$$+f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\left(\xi^{m,J}\cdot H^{mJ,j}+d^{mJ}\cdot\varphi^{m,J,j}+\xi^{mJ}\cdot\varphi^{m,J,j}\right)$$

$$+C_{12}\left(d^{(m+1)J}\cdot H^{(m+1)J,j}-d^{mJ}\cdot H^{mJ,j}\right)^2$$

$$=f_j\left(d^{mJ}\cdot H^{mJ,j}\right)-\frac{1}{\eta_m}\Delta_j d^{mJ}\cdot\xi^{m,J}-\frac{1}{\eta_m}\sum_{i=1}^{n}\left(\Delta_j b_i^{mJ}\cdot\tau_i^{m,J}\right)$$

$$+f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\sum_{i=1}^{n}d_i^{mJ}\left(\tau_i^{m,J}\cdot x^j\right)^2\cdot\int_0^1(1-t)h''\left(b_i^{mJ}\cdot g\left(v_i^{mJ}x^j\right)+t\left(\tau_i^{mJ}\cdot x^j\right)\right)dt$$

$$+f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\xi^{m,J}\cdot\varphi^{m,J,j}+C_{12}\left(d^{(m+1)J}\cdot H^{(m+1)J,j}-d^{mJ}\cdot H^{mJ,j}\right)^2. \tag{68}$$

Summing (68) from $j=1$ to $j=J$ and noting (4) - (7), (37) and (39), we have

$$E\left(w^{(m+1)J}\right)\le E\left(w^{mJ}\right)$$

$$-\frac{1}{\eta_m}\left(\left\|\sum_{j=1}^{J}\Delta_j d^{mJ}\right\|^2+\sum_{i=1}^{p}\left\|\sum_{j=1}^{J}\Delta_j b_i^{mJ}\right\|^2+\sum_{i=1}^{n}\left\|\sum_{j=1}^{J}\Delta_j v_i^{mJ}\right\|^2\right)+\delta_m$$

$$=E\left(w^{mJ}\right)-\eta_m\left(\left\|E_d\left(w^{mJ}\right)\right\|^2+\sum_{i=1}^{p}\left\|E_{b_i}\left(w^{mJ}\right)\right\|^2+\sum_{i=1}^{n}\left\|E_{v_i}\left(w^{mJ}\right)\right\|^2\right)+\delta_m$$

$$=E\left(w^{mJ}\right)-\eta_m\left\|E_w\left(w^{mJ}\right)\right\|^2+\delta_m, \tag{69}$$

where

$$\delta_m=-\frac{1}{\eta_m}\sum_{j=1}^{J}\Delta_j d^{mJ}\cdot\sum_{j=1}^{J}R^{m,j}-\frac{1}{\eta_m}\sum_{i=1}^{n}\left(\sum_{j=1}^{J}\Delta_j b_i^{mJ}\cdot\sum_{j=1}^{J}\Delta_j Q_i^{mJ}\right)$$

$$-\frac{1}{\eta_m}\sum_{i=1}^{n}\left(\sum_{j=1}^{J}\Delta_j v_i^{mJ}\cdot\sum_{j=1}^{J}\Delta_j r_i^{mJ}\right)$$

$$+\sum_{j=1}^{J}\sum_{i=1}^{n}d_i^{mJ}f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\left(\tau_i^{m,J}\cdot x^j\right)^2\cdot$$

$$\times\int_0^1(1-t)\,h''\left(b_i^{mJ}\cdot g\left(v_i^{mJ}x^j\right)+t\left(\tau_i^{mJ}\cdot x^j\right)\right)dt$$

$$+\sum_{j=1}^{J}f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\xi^{m,J}\cdot\varphi^{m,J,j}+C_{12}\big(d^{(m+1)J}\cdot H^{(m+1)J,j}-d^{mJ}\cdot H^{mJ,j}\big)^2.$$

It now follows from (45) and (46) that

$$\left\|H^{mJ+j,k}\right\|=\left\|H\left(B^{mJ+j}\cdot G\left(V^{mJ+j}x^k\right)\right)\right\|\le C_4,$$

$$\left|d^{mJ+j}\cdot H^{mJ+j,k}\right|\le\left\|d^{mJ+j}\right\|\left\|H^{mJ+j,k}\right\|\le C_2C_4\equiv D_3. \tag{70}$$

By (16), (51)-(53) and (64), the first term of δ_m can be estimated as follows.

$$\left\|-\frac{1}{\eta_m}\sum_{j=1}^{J}\Delta_j d^{mJ}\cdot\sum_{j=1}^{J}R^{m,j}\right\|\le\frac{1}{\eta_m}\sum_{j=1}^{J}\left\|\Delta_j d^{mJ}\right\|\cdot\sum_{j=1}^{J}\left\|R^{m,j}\right\|\le C_{11,1}\eta_m^2, \tag{71}$$

where $C_{11,1}=J^2C_4C_{5,1}C_8=JC_5C_8$.

Similar estimates for the other terms of δ_m can be obtained with corresponding constants $C_{11,t}>0$ for $t=2,\ldots,5$. Finally, the desired estimate (66) is proved by setting $C_{11}=\sum_{t=1}^{5}C_{11,t}$. □

Now, we are ready to prove the convergence theorem.

4.2 Proof of theorem 1 for OGM-F.

The proof is divided into two parts, dealing with (28) and (29), respectively.

Proof. (*for* (28)) By (A2) and Lemma 5, we conclude that

$$\sum_{m=0}^{\infty}\eta_m\left\|E_w\left(w^{mJ}\right)\right\|^2=\sum_{m=0}^{\infty}\eta_m\left(\left\|E_d\left(w^{mJ}\right)\right\|^2+\left\|E_B\left(w^{mJ}\right)\right\|^2+\left\|E_V\left(w^{mJ}\right)\right\|^2\right)<\infty, \tag{72}$$

$$\sum_{m=0}^{\infty}\eta_m\left\|E_d\left(w^{mJ}\right)\right\|^2<\infty. \tag{73}$$

Employing the integral Taylor expansion, we deduce that

$$f'_j\left(d^{(m+1)J}\cdot H^{(m+1)J,j}\right)H^{(m+1)J,j}-f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)H^{mJ,j}$$

$$=f'_j\left(d^{(m+1)J}\cdot H^{(m+1)J,j}\right)\varphi^{m,J,j}$$

$$+\left(f'_j\left(d^{(m+1)J}\cdot H^{(m+1)J,j}\right)-f'_j\left(d^{mJ}\cdot H^{(m+1)J,j}\right)\right)H^{mJ,j}$$
$$+\left(f'_j\left(d^{mJ}\cdot H^{(m+1)J,j}\right)-f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\right)H^{mJ,j}$$
$$=f'_j\left(d^{(m+1)J}\cdot H^{(m+1)J,j}\right)\varphi^{m,J,j}$$
$$+\left(\xi^{m,J}\cdot H^{(m+1)J,j}\right)H^{mJ,j}\cdot\int_0^1(1-t)f''_j\left(d^{mJ}\cdot H^{(m+1)J,j}+t\left(\xi^{mJ}\cdot H^{(m+1)J,j}\right)\right)dt$$
$$+\left(d^{mJ}\cdot\varphi^{m,J,j}\right)H^{mJ,j}\cdot\int_0^1(1-t)f''_j\left(d^{mJ}\cdot H^{mJ,j}+t\left(d^{mJ}\cdot\varphi^{m,J,j}\right)\right)dt \tag{74}$$

Note (A2) and let $\eta_c>0$ be an upper bound of $\{\eta_m\}_{m=0}^{\infty}$. It follows from (45)-(47) that

$$\left|d^{mJ}\cdot H^{(m+1)J,j}+t\left(\xi^{m,J}\cdot H^{(m+1)J,j}\right)\right|\le\left(\left\|d^{mJ}\right\|+\left\|\xi^{m,J}\right\|\right)\left\|H^{(m+1)J,j}\right\|$$
$$\le\left(C_2+C_5\eta_c\right)C_4,\quad t\in(0,\,1)\,, \tag{75}$$
$$\left|d^{mJ}\cdot H^{mJ,j}+t\left(d^{m,J}\cdot\varphi^{m,J,j}\right)\right|\le\left\|d^{mJ}\right\|\left(\left\|H^{mJ,j}\right\|+\left\|\varphi^{m,J,j}\right\|\right)$$
$$\le C_2\left(C_4+C_7\eta_c\right)=D_3+C_2C_7\eta_c,\quad t\in(0,\,1)\,. \tag{76}$$

According to (75), (76) and the proof of Lemma 1, there are positive constants $C_{10},\ C_{11}>0$ such that

$$\left|\int_0^1(1-t)\,f''_j\left(d^{mJ}\cdot H^{(m+1)J,j}+t\xi^{m,J}\cdot H^{(m+1)J,j}\right)dt\right|\le C_{13}, \tag{77}$$

$$\left|\int_0^1(1-t)\,f''_j\left(d^{mJ}\cdot H^{mJ,j}+t\xi^{m,J}\cdot\varphi^{m,J,j}\right)dt\right|\le C_{14}. \tag{78}$$

By (54), we obtain $\left|d^{(m+1)J}\cdot H^{(m+1)J,j}\right|\le C_2C_4=D_3$. Employing (5) and (46), (47), (56), (77) and (78), and summing (74) from $j=1$ to $j=J$, we conclude that

$$\left|\left\|E_u\left(w^{(m+1)J}\right)\right\|-\left\|E_u\left(w^{mJ}\right)\right\|\right|\le\left\|E_u\left(w^{(m+1)J}\right)-E_u\left(w^{mJ}\right)\right\|$$
$$\le C_{13}J\max_{\substack{1\le j\le J\\ m\in\mathbb{N}}}\left(\left\|H^{(m+1)J,j}\right\|\left\|H^{mJ,j}\right\|\right)\left\|\xi^{m,J}\right\|$$
$$+\left(JC_{5,1}+C_{14}J\max_{\substack{1\le j\le J\\ m\in\mathbb{N}}}\left(\left\|d^{mJ}\right\|\left\|H^{mJ,j}\right\|\right)\right)\max_{\substack{1\le j\le J\\ m\in\mathbb{N}}}\left(\left\|\varphi^{m,J,j}\right\|\right)$$
$$\le C_{15}\eta_m, \tag{79}$$

where $C_{15}=JC_4^2C_5C_{13}+JC_{5,1}C_7+JD_3C_7C_{14}$. Combining (72), (79) and Lemma 2 results in $\lim\limits_{m\to\infty}\left\|E_d\left(w^{mJ}\right)\right\|=0$.

Similarly as in the proof to (79), there exists a positive constant C_{16} such that

$$\left\|E_d\left(w^{mJ+j}\right)-E_d\left(w^{mJ}\right)\right\|\le C_{16}\eta_m. \tag{80}$$

Since

$$\begin{aligned}\left\|E_d\left(w^{mJ+j}\right)\right\| &\leq \left\|E_d\left(w^{mJ+j}\right)-E_d\left(w^{mJ}\right)\right\|+\left\|E_d\left(w^{mJ}\right)\right\| \\ &\leq C_{16}\eta_m+\left\|E_d\left(w^{mJ}\right)\right\|, \end{aligned} \tag{81}$$

we have $\lim\limits_{m\to\infty}\left\|E_d\left(w^{mJ+j}\right)\right\| = 0$ for $j = 1,\ 2,\ldots,J$. Similarly, we deduce that $\lim\limits_{m\to\infty}\left\|E_{b_i}\left(w^{mJ+j}\right)\right\| = 0$, $\lim\limits_{m\to\infty}\left\|E_{v_i}\left(w^{mJ+j}\right)\right\| = 0$ for $n = 1,\ldots,n,\ \ j = 1,\ 2,\ldots,J$, and

$$\lim_{m\to\infty}\left\|E_w\left(w^{mJ+j}\right)\right\| = 0, \quad j = 1,\ 2,\ldots,J. \tag{82}$$

This immediately gives

$$\lim_{m\to\infty}\left\|E_w\left(w^m\right)\right\| = 0. \tag{83}$$

□

Proof. (*for* (29)) According to (A3), the sequence $\{w^m\}$ $(m\in\mathbb{N})$ has a subsequence $\{w^{m_k}\}$ $(k\in\mathbb{N})$ that is convergent to, say, $w^*\in\Omega_0$. It follows from (28) and the continuity of $E_w(w)$ that

$$\left\|E_w\left(w^*\right)\right\| = \lim_{m\to\infty}\left\|E_w\left(w^{m_k}\right)\right\| = \lim_{m\to\infty}\left\|E_w\left(w^m\right)\right\| = 0. \tag{84}$$

This implies that w^* is a stationary point of $E(w)$. Hence, $\{w^m\}$ has at least one accumulation point and every accumulation point must be a stationary point.

Next, we prove that $\{w^m\}$ has precisely one accumulation point. Let us assume to the contrary that $\{w^m\}$ has at least two accumulation points $\bar{w}\neq\tilde{w}$. We write $w^m = \left(w_1^m, w_2^m,\ldots,w_{p(s+n)+n}^m\right)^T$. It is easy to see from (12)-(14) that $\lim_{m\to\infty}\left\|w^{m+1}-w^m\right\| = 0$, or equivalently, $\lim_{m\to\infty}\left|w_i^{m+1}-w_i^m\right| = 0$ for $i = 1,\ 2,\ldots,p(s+n)+n$. Without loss of generality, we assume that the first components of $\bar{w}$ and $\tilde{w}$ do not equal to each other, that is, $\bar{w}_1\neq\tilde{w}_1$. For any real number $\lambda\in(0,\ 1)$, $w_1^\lambda = \lambda\bar{w}_1+(1-\lambda)\tilde{w}_1$. By Lemma 3, there exists a subsequence $\left\{w_1^{m_{k_1}}\right\}$ of $\{w_1^m\}$ converging to w_1^λ as $k_1\to\infty$. Due to the boundedness of $\left\{w_2^{m_{k_1}}\right\}$, there is a convergent subsequence $\left\{w_2^{m_{k_2}}\right\}\subset\left\{w_2^{m_{k_1}}\right\}$. We define $w_2^\lambda = \lim_{k_2\to\infty}w_2^{m_{k_2}}$. Repeating this procedure, we end up with decreasing subsequences $\{m_{k_1}\}\supset\{m_{k_1}\}\supset\cdots\supset\left\{m_{k_{p(s+n)+n}}\right\}$ with $w_i^\lambda = \lim_{k_i\to\infty}w_2^{m_{k_i}}$ for each $i = 1,\ 2,\ldots,p(s+n)+n$. Write $w^\lambda = \left(w_1^\lambda, w_2^\lambda,\ldots,w_{p(s+n)+n}^\lambda\right)^T$. Then, we see that w^λ is an accumulation point of $\{w^m\}$ for any $\lambda\in(0,\ 1)$. But this means that $\Omega_{0,1}$ has interior points, which contradicts (A4). Thus, w^* must be a unique accumulation point of $\{w^m\}_{m=0}^\infty$. This completes the proof of the strong convergence. □

4.3 Convergence analysis for OGM-SS approach

Now, let the sequence $\left\{w^{mJ+j}\right\}$ $(m\in\mathbb{N},\ \ j = 1,\ 2,\ldots,J)$ be generated by (19), (20) and (21), and let

$$R^{m,j} = \Delta_j^m d^{mJ+j}-\Delta_j^m d^{mJ}, \tag{85}$$

$$Q_i^{m,j} = \Delta_j^m b_i^{mJ+j} - \Delta_j^m b_i^{mJ}, \tag{86}$$

$$r_i^{m,j} = \Delta_j^m v_i^{mJ+j} - \Delta_j^m v_i^{mJ}, \tag{87}$$

$$\xi^{m,l} = d^{mJ+l} - d^{mJ} = \sum_{j=1}^{l} \Delta_j^m d^{mJ+j} = \sum_{j=1}^{l} \Delta_j^m d^{mJ} + \sum_{j=1}^{l} R^{m,j}, \tag{88}$$

$$\tau_i^{m,l} = b_i^{mJ+l} - b_i^{mJ} = \sum_{j=1}^{l} \Delta_j^m b_i^{mJ+j} = \sum_{j=1}^{l} \Delta_j^m b_i^{mJ} + \sum_{j=1}^{l} Q_i^{m,j}, \tag{89}$$

$$\phi_i^{m,l} = v_i^{mJ+l} - v_i^{mJ} = \sum_{j=1}^{l} \Delta_j^m v_i^{mJ+j} = \sum_{j=1}^{l} \Delta_j^m v_i^{mJ} + \sum_{j=1}^{l} r_i^{m,j}, \tag{90}$$

$$\psi^{m,l,j} = G^{mJ+l,m,j} - G^{mJ,m,j}, \tag{91}$$

$$\varphi^{m,l,j} = H^{mJ+l,m,j} - H^{mJ,m,j}, \tag{92}$$

$$m \in \mathbb{N};\ \ j = 1, 2, \ldots, J;\ \ l = 1, 2, \ldots, J,\ \ i = 1, 2, \ldots, n.$$

It is obvious that Lemmas 1-3 are not influenced by the new definitions. In place of Lemmas 4 and 5, we now have the following two Lemmas.

Lemma 6. *Let conditions (A1) and (A3) be valid, and let the sequence* $\left\{w^{mJ+j}\right\}$ *be generated by* (19), (20) *and* (21). *Then there are constants* $C_3 - C_{10}$ *such that*

$$\left\|G^{mJ+j,m,k}\right\| \le C_3,\ \ \left\|H^{mJ+j,m,k}\right\| \le C_4, \tag{93}$$

$$\left\|\xi^{m,l}\right\| \le C_5\eta_m,\ \ \left\|\psi^{m,l,j}\right\| \le C_6\eta_m,\ \ \left\|\varphi^{m,l,j}\right\| \le C_7\eta_m, \tag{94}$$

$$\left\|R^{m,l}\right\| \le C_8\eta_m^2,\ \ \left\|Q_i^{m,j}\right\| \le C_9\eta_m^2,\ \ \left\|r_i^{m,j}\right\| \le C_{10}\eta_m^2, \tag{95}$$

where $m \in \mathbb{N};\ \ j,\, k = 1, 2, \ldots, J;\ \ l = 1, 2, \ldots, J,\ \ i = 1, 2, \ldots, n.$

Proof. According to (45), we have

$$\left|v_i^{mJ+j} \cdot x^{m,k}\right| \le \left\|v_i^{mJ+j}\right\| \left\|x^k\right\| \le C_1C_2 \equiv D_1. \tag{96}$$

Thus, there exists a positive constant $C_{3,1}$ such that

$$\max_{|t|\le D_1} |g(t)| = C_{3,1}, \tag{97}$$

$$\left\|G^{mJ+j,m,k}\right\| = \left\|G\left(V^{mJ+j}x^{m,k}\right)\right\| \le \sqrt{n}C_{3,1} \equiv C_3. \tag{98}$$

Similarly, (94) and (95) can be proved after adjusting the corresponding superscripts in the proof to Lemma 4. □

Lemma 7. *Let the sequence* $\left\{w^{mJ+j}\right\}$ *be generated by* (19), (20) *and* (21). *Under assumptions (A1) and (A3), there holds*

$$E\left(w^{(m+1)J}\right) \le E\left(w^{mJ}\right) - \eta_m\left\|E_w\left(w^{mJ}\right)\right\|^2 + C_{11}\eta_m^2,\ \ (m = 0,\, 1, \ldots) \tag{99}$$

where $C_{11} > 0$ *is a constant defined in Lemma 5*

Proof. As in the proof to Lemma 6, we only need to adjust some superscripts. For example, corresponding to (67), we change the related superscripts and get

$$f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)d^{mJ}\cdot\varphi^{m,J,j}=f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\sum_{i=1}^{n}d_i^{mJ}h'\left(b_i^{mJ}\cdot g\left(v_i^{mJ}\cdot x^{m,j}\right)\right)\tau_i^{m,J}\cdot x^{m,j}$$

$$+f'_j\left(d^{mJ}\cdot H^{mJ,j}\right)\sum_{i=1}^{n}d_i^{mJ}\left(\tau_i^{m,l}\cdot x^{m,j}\right)^2$$

$$\times\int_0^1(1-t)h''\left(b_i^{mJ}\cdot g\left(v_i^{mJ}x^{m,j}\right)+t\left(\tau_i^{mJ}\cdot x^{m,j}\right)\right)dt. \tag{100}$$

The details of prove are left to the interested users. □

Proof of theorem 1 for OGM-SS approach. To get the prove of OGM-SS approache we may use Lemmas 1-3 and Lemmas 6-7 results as in the proof to Theorem 1 for OGM-F.

5 Conclusion

In this paper, we show a some comprehensive results on the weak and strong convergence for four layer backpropagation feedforward neural networks. The assumptions in this paper are much more relaxed comparing with the other existing convergence results. This convergence analysis holds to implement multilayered neural networks with various kinds of activation fucntions in hidden layers.

References

[1] Baldi P., Vershynin R. The capacity of feedforward neural networks. Neural Networks, Vol. 116, 288–311 (2019).

[2] Ismailov V.E. Approximation by neural networks with weights varying on a finite set of directions. Journal of Mathematical Analysis and Applications, Vol. 389, Issue 1, 72–83 (2012).

[3] Chen Z., Cao F. The approximation operators with sigmoidal functions. Computers and Mathematics with Applications, Vol. 58, Issue 4, 758–765 (2009).

[4] Pinkus A. Approximation theory of the MLP model in neural networks. Acta Numerica, Vol. 8, 143-195 (1999).

[5] Yamashita R., Nishio M., Do R.K., Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging, Vol. 9, Issue 4, 611–629 (2018).

[6] Werbos P.J. Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis. Harvard University, Cambridge, MA, (1974).

[7] Marakhimov A.R., Khudaybergenov K.K. A Fuzzy MLP Approach for Identification of Nonlinear Systems. Contemporary Mathematics. Fundamental Directions, Vol. 65, Issue 1, 44–53 (2019).

[8] LeCun Y. Une procedure d'apprentissage pour reseau a seuil asymmetrique. A la Frontieredel'Intelligence Ariticielle des Sciences de la Connaissance des Neurosciences, Vol. 85, 599–604 (1985).

[9] Rumelhart D.E., Hinton G.E., Williams R.J. Learning representations by back-propagation errors. Nature, Vol. 323, 533–536 (1986).

[10] Wilson D.R., Martinez T. R. The general inefficiency of batch training for gradient descent learning. Neural Networks, Vol. 16, 1429–1451 (2003).

[11] Nakama T. Theoretical analysis of batch and on-line training for gradient descent learning in neural networks. Neurocomputing, Vol. 73, 151–159 (2009).

[12] Zhang H.S., Wu W., Liu F., Yao M.C. Boundedness and convergence of online gradient method with penalty for feedforward neural networks. IEEE Transactions on Neural Networks, Vol. 20, 1050–1054 (2009).

[13] Bertsekas D.P., Tsitsiklis J.N. Neuro-dynamic programming. Athena Scientific, (1996).

[14] Wu W., Feng G.R., Li X. Training multilayer perceptrons via minimization of sum of ridge functions. Advances in Computational Mathematics, Vol. 17, 331–347 (2002).

[15] Wu W., Shao H.M., Qu D. Strong convergence of gradient methods for BP networks training. In Proceedings of 2005 international conference on neural networks and brains, 332–334 (2005).

[16] Xu Z.B., Zhang R., Jin W.F. When on-line BP training converges. IEEE Transactions on Neural Networks, Vol. 20, 1529–1539 (2009).

[17] Wua W., Wanga J., Chenga M., Li Z. Convergence analysis of online gradient method for BP neural networks. Neural Networks, Vol. 24, 91–98 (2011).