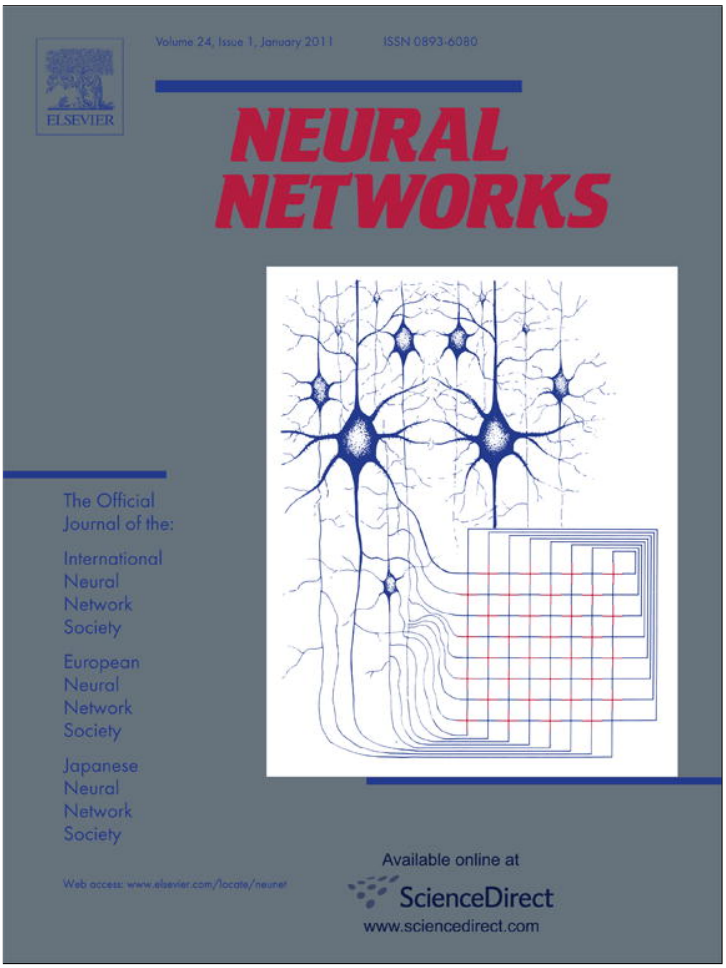

Neural Networks 24 (2011) 91–98

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Convergence analysis of online gradient method for BP neural networks☆

Wei Wu [a,*], Jian Wang [a,b], Mingsong Cheng [a], Zhengxue Li [a]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, PR China

[b] School of Mathematics and Computational Sciences, Petroleum University of China, Dongying, 257061, PR China

ARTICLE INFO

Article history:
Received 28 March 2010
Received in revised form 8 September 2010
Accepted 9 September 2010

Keywords:
Neural networks
Backpropagation learning
Online gradient method
Weak convergence
Strong convergence

ABSTRACT

This paper considers a class of online gradient learning methods for backpropagation (BP) neural networks with a single hidden layer. We assume that in each training cycle, each sample in the training set is supplied in a stochastic order to the network exactly once. It is interesting that these stochastic learning methods can be shown to be deterministically convergent. This paper presents some weak and strong convergence results for the learning methods, indicating that the gradient of the error function goes to zero and the weight sequence goes to a fixed point, respectively. The conditions on the activation function and the learning rate to guarantee the convergence are relaxed compared with the existing results. Our convergence results are valid for not only S–S type neural networks (both the output and hidden neurons are Sigmoid functions), but also for P–P, P–S and S–P type neural networks, where S and P represent Sigmoid and polynomial functions, respectively.

1. Introduction

Artificial neural network has been a hot topic in recent years in cognitive science, computational intelligence and intelligent information processing. Backpropagation (BP) is the most broadly used learning method for feedforward neural networks. It was first proposed by Werbos (1974) in his Ph.D. thesis, and has been rediscovered several times (LeCun, 1985; Parker, 1982; Rumelhart, Hinton, & Williams, 1986). There are two practical ways to implement the backpropagation algorithm: batch updating approach and online updating approach. Corresponding to the standard gradient method, the batch updating approach accumulates the weight correction over all the training samples before actually performing the update. On the other hand, the online updating approach updates the network weights immediately after each training sample is fed. Some authors compare the two different training schemes for feedforward neural networks (Heskes & Wiegerinck, 1996; Nakama, 2009; Wilson & Martinez, 2003). Heskes and Wiegerinck (1996) reveal several asymptotic properties of the two schemes. Wilson and Martinez (2003) explain why batch training is almost always slower than online training (often orders of magnitude slower) especially on large training sets. Nakama (2009) theoretically analyzes the convergence properties of the two schemes applied to quadratic loss functions and shows the exact degrees to which the training set size, the variance of the per-instance gradient, and the learning rate affect the rate of convergence for each scheme.

There are three approaches for online training of BP neural networks according to different fashions of sampling. The first approach is OGM-CS (completely stochastic order): At each learning step, one of the samples is drawn at random from the training set and presented to the network (Finnoff, 1994; Heskes & Wiegerinck, 1996; Terence, 1989; Wilson & Martinez, 2003). The second approach is OGM-SS (special stochastic order): In each training cycle, each sample in the training set is supplied in a stochastic order to the network exactly once (Heskes & Wiegerinck, 1996; Li & Ding, 2005; Li, Wu, & Tian, 2004; Nakama, 2009). The third approach is OGM-F (fixed order): In each training cycle, each sample in the training set is supplied in a fixed order to the network exactly once (Heskes & Wiegerinck, 1996; Mangasarian & Solodov, 1994; Wu & Xu, 2002; Wu, Feng, Li, & Xu, 2005; Xu, Zhang, & Jin, 2009).

Naturally, the existing convergence results for OGM-CS are mostly asymptotic convergence with a probabilistic nature as the size of training samples goes to infinity (Bertsekas & Tsitsiklis, 1996; Chakraborty & Pal, 2003; Fine & Mukherjee, 1999; Finnoff, 1994; Liang, Feng, Lee, Lim, & Lee, 2002; Tadic & Stankovic, 2000; Terence, 1989; Zhang, Wu, Liu, & Yao, 2009). Deterministic convergence can be obtained for OGM-SS and OGM-F (Li et al., 2004; Mangasarian & Solodov, 1994; Shao, Wu, & Liu, 2007; Wu & Xu, 2002; Wu et al., 2005; Wu, Feng, & Li, 2002; Wu & Shao, 2003; Wu, Shao, & Qu, 2005; Xu et al., 2009). It is interesting to see that the learning method OGM-SS with stochastic nature enjoys deterministic convergence. The convergence result is a

☆ Project supported by the National Natural Science Foundation of China (No. 10871220).

* Corresponding author.
E-mail address: wuweiw@dlut.edu.cn (W. Wu).

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.09.007

bit easier to prove for OGM-F than for OGM-SS. But we have reason to believe, and our experience shows, that OGM-SS behaves numerically better than OGM-F since the stochastic nature of the learning procedure survives in OGM-SS (Li & Ding, 2005; Li et al., 2004).

To guarantee the convergence, it is commonly required that the learning rate η_m satisfies the assumptions $\sum_{m=1}^{\infty}\eta_m = \infty$ and $\sum_{m=1}^{\infty}\eta_m^2 < \infty$ as in Bertsekas and Tsitsiklis (1996) and Tadic and Stankovic (2000) for OGM-CS. An extra assumption $\lim_{m\to\infty}\eta_m/\eta_{m+1} = 1$ was introduced by Xu et al. (2009) for OGM-F. A special condition which is basically $\eta_m = O(1/m)$ was required in Li et al. (2004), Shao et al. (2007), Wu and Xu (2002), Wu et al. (2005), Wu et al. (2002), Wu and Shao (2003) and Wu et al. (2005) for OGM-F and OGM-SS.

To obtain the strong convergence result, which means that the weight sequence converges to a fixed point, Wu et al. (2005) introduced an additional assumption: the number of the stationary points of the error function is finite. A more relaxed condition is used in Xu et al. (2009): the gradient of the error function has at most countably infinite number of stationary points.

The aim of this paper is to present a comprehensive study on the weak and strong convergence for OGM-F and OGM-SS, indicating that the gradient of the error function goes to zero and the weight sequence goes to a fixed point, respectively. These convergence results improve the existing results in Li et al. (2004), Shao et al. (2007), Wu and Xu (2002), Wu et al. (2005), Wu et al. (2002), Wu and Shao (2003), Wu et al. (2005) and Xu et al. (2009) such that the conditions on the activation function and the learning rate to guarantee the convergence are much relaxed. Specifically, we make the following contributions:

- The extra condition $\lim_{m\to\infty}\eta_m/\eta_{m+1} = 1$ for the learning rate is removed which is a requisite in Xu et al. (2009).
- The convergence results are valid for both OGM-F and OGM-SS.
- The convergence results apply not only to S–S type neural networks (both the output and hidden neurons are Sigmoid functions), but also to P–P, P–S and S–P type neural networks, where S and P represent Sigmoid and polynomial functions, respectively.
- The restrictive assumptions for the strong convergence in Wu et al. (2005) and Xu et al. (2009) are relaxed such that the stationary points set of the error function is only required not to contain any interior point.
- We assume that the derivative g' of the activation function is Lipschitz continuous on any bounded closed interval. This improves the corresponding conditions in Wu et al. (2005), which require the boundedness of the second derivative g'', and in Xu et al. (2009), which require g' to be Lipschitz continuous and uniformly bounded on the whole R.

Let us make a few remarks on the above contribution points. For the first contribution point, as an example, we recall a well-known adaptive technique for the learning rate η_m: $\eta_m = (1+a)\eta_{m-1}$ if the error is decreasing, and $\eta_m = (1-a)\eta_{m-1}$ otherwise, where $a < 1$ is a positive number. Xu's condition $\lim_{m\to\infty}\eta_m/\eta_{m+1} = 1$ (Xu et al., 2009) is not valid in this case, while our convergence results remains valid. For the second contribution point, it is interesting to see that the learning method OGM-SS with stochastic nature enjoys deterministic convergence. We observe that OGM-F is actually a deterministic iteration procedure in that the iteration sequence is determined uniquely by the initial value and the fixed order of the samples. The convergence result is a bit easier to prove for OGM-F than for OGM-SS. We have reason to believe, and our experience shows, that OGM-SS behaves numerically better than OGM-F since the stochastic nature of the learning procedure survives in OGM-SS (Li & Ding, 2005; Li et al., 2004). Our convergence results are generalizations of both the results of Xu et al. (2009), which considers OGM-F, and the results of Li et al. (2004), which considers OGM-SS with an unpleasant condition $\eta_m = O(1/m)$ on the learning rate. Our third contribution allows the activation functions for both hidden and output layers to be more flexible. Here we remark that typically, S–S type networks are used for classification problems, and S–P type networks with Sigmoid hidden neurons and linear output neurons are used for approximation problems. The existing convergence results (Li et al., 2004; Shao et al., 2007; Wu & Xu, 2002; Wu et al., 2005, 2002; Wu & Shao, 2003; Wu et al., 2005; Xu et al., 2009) are mostly for either S–S type or S–P type alone but not for both of them. In this paper, we give a uniform treatment for all types of networks. The fourth and fifth contribution points are mainly of theoretical interest. From a theoretical point of view, we mention that different analytical tools are employed in Wu et al. (2005) and Xu et al. (2009) and this study for the convergence analysis, might explain, at least in part, why different conditions are obtained for the convergence. The differential Taylor expansion is used in Wu et al. (2005), which requires the boundedness of the second derivative g'' of the activation function g; the mean value theorem of integrals is employed in Xu et al. (2009), which requires g' to be Lipschitz continuous and uniformly bounded; and in this paper, we use the integral Taylor expansion and hence require the Lipschitz continuity of g' on any bounded closed interval. Finally, we point out that Xu et al. (2009) is a big step forward for the convergence study of OGM-F and that Xu et al. (2009) also includes another convergence result under the condition that the error function is directionally convex. This convex condition is not considered in this paper.

The rest of this paper is organized as follows. In Section 2, online updating methods including OGM-F and OGM-SS are introduced. The main convergence results are presented in Section 3 and their proofs are gathered in Section 4. Some conclusions are drawn in Section 5.

2. OGM-F and OGM-SS

Let us begin with an introduction of a feedforward neural network with three layers. The numbers of neurons for the input, hidden and output layers are p, n and 1, respectively. Suppose that the training sample set is $\left\{\mathbf{x}^j, O^j\right\}_{j=1}^{J} \subset \mathbb{R}^p \times \mathbb{R}$, where $\mathbf{x}^j$ and O^j are the input and the corresponding ideal output of the jth sample, respectively. Let $\mathbf{V} = \left(v_{i,j}\right)_{n\times p}$ be the weight matrix connecting the input and the hidden layers, and write $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{ip})^T$ for $i = 1, 2, \ldots, n$. The weight vector connecting the hidden and the output layers is denoted by $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T \in \mathbb{R}^n$. To simplify the presentation, we combine the weight matrix $\mathbf{V}$ with the weight vector $\mathbf{u}$, and write $\mathbf{w} = \left(\mathbf{u}^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_n^T\right)^T \in \mathbb{R}^{n(p+1)}$. Let $g, f : \mathbb{R} \to \mathbb{R}$ be given activation functions for the hidden and output layers, respectively. For convenience, we introduce the following vector valued function

$$G(\mathbf{z}) = (g(z_1), g(z_2), \ldots, g(z_n))^T, \quad \forall\, \mathbf{z} \in \mathbb{R}^n. \tag{1}$$

For any given input $\mathbf{x} \in \mathbb{R}^p$, the output of the hidden neurons is $G(\mathbf{Vx})$, and the final actual output is

$$y = f(\mathbf{u} \cdot G(\mathbf{Vx})). \tag{2}$$

For any fixed weights $\mathbf{w}$, the error of the neural networks is defined as

$$E(\mathbf{w}) = \frac{1}{2}\sum_{j=1}^{J}(O^j - f(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j)))^2 = \sum_{j=1}^{J} f_j(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j)), \tag{3}$$

where $f_j(t) = \frac{1}{2}(O^j - f(t))^2$, $j = 1, 2, \ldots, J$, $t \in \mathbb{R}$. The gradients of the error function with respect to $\mathbf{u}$ and $\mathbf{v}_i$ are, respectively,

given by

$$E_{\mathbf{u}}(\mathbf{w}) = -\sum_{j=1}^{J}\left(O^j - y^j\right) f'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))G(\mathbf{V}\mathbf{x}^j) = \sum_{j=1}^{J} f_j'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))G(\mathbf{V}\mathbf{x}^j), \tag{4}$$

$$E_{\mathbf{v}_i}(\mathbf{w}) = -\sum_{j=1}^{J}\left(O^j - y^j\right) f'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))u_i g'(\mathbf{v}_i\cdot\mathbf{x}^j)\mathbf{x}^j = \sum_{j=1}^{J} f_j'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))u_i g'(\mathbf{v}_i\cdot\mathbf{x}^j)\mathbf{x}^j, \tag{5}$$

where

$$y^j = f(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j)), \quad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, J. \tag{6}$$

Write

$$E_{\mathbf{V}}(\mathbf{w}) = \left(E_{\mathbf{v}_1}(\mathbf{w})^T, E_{\mathbf{v}_2}(\mathbf{w})^T, \ldots, E_{\mathbf{v}_n}(\mathbf{w})^T\right)^T, \tag{7}$$

$$E_{\mathbf{w}}(\mathbf{w}) = \left(E_{\mathbf{u}}(\mathbf{w})^T, E_{\mathbf{V}}(\mathbf{w})^T\right)^T. \tag{8}$$

First, let us consider the case that the training samples are supplied to the network in a fixed order (OGM-F) in the training process. Hence, starting from an arbitrary initial guess $\mathbf{w}^0$, we proceed to refine it iteratively by the formulas

$$\mathbf{u}^{mJ+j+1} = \mathbf{u}^{mJ+j} + \Delta_j\mathbf{u}^{mJ+j}, \tag{9}$$

$$\mathbf{v}_i^{mJ+j+1} = \mathbf{v}_i^{mJ+j} + \Delta_j\mathbf{v}_i^{mJ+j}, \tag{10}$$

where

$$\Delta_k\mathbf{u}^{mJ+j} = \eta_m(O^k - y^{mJ+j,\,k})f'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,k})G^{mJ+j,\,k} = -\eta_m f_k'\left(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,k}\right)G^{mJ+j,\,k}, \tag{11}$$

$$\Delta_k\mathbf{v}_i^{mJ+j} = \eta_m\left(O^k - y^{mJ+j,\,k}\right)f'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,k}) \times u_i^{mJ+j}g'(\mathbf{v}_i^{mJ+j}\cdot\mathbf{x}^k)\mathbf{x}^k = -\eta_m f_k'\left(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,k}\right)u_i^{mJ+j}g'(\mathbf{v}_i^{mJ+j}\cdot\mathbf{x}^k)\mathbf{x}^k, \tag{12}$$

$$G^{mJ+j,\,k} = G(\mathbf{V}^{mJ+j}\mathbf{x}^k), \qquad y^{mJ+j,\,k} = f(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,k}), \quad m\in\mathbb{N};\ i = 1, 2, \ldots, n;\ j,\ k = 1, 2, \ldots, J. \tag{13}$$

Here the parameter η_m is the learning rate, whose value may be changed after each cycle of the training procedure.

We can also choose training samples in a special stochastic order (OGM-SS) as follows: For the mth training cycle, let $\left\{\mathbf{x}^{m,\,1}, \mathbf{x}^{m,\,2}, \ldots, \mathbf{x}^{m,\,J}\right\}$ be a stochastic permutation of the set $\left\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^J\right\}$. Similar to (9) and (10), the weights are iteratively updated in the following fashion

$$\mathbf{u}^{mJ+j+1} = \mathbf{u}^{mJ+j} + \Delta_j^m\mathbf{u}^{mJ+j}, \tag{14}$$

$$\mathbf{v}_i^{mJ+j+1} = \mathbf{v}_i^{mJ+j} + \Delta_j^m\mathbf{v}_i^{mJ+j}, \tag{15}$$

where

$$\Delta_k^m\mathbf{u}^{mJ+j} = \eta_m(O^k - y^{mJ+j,\,m,\,k})f'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,m,\,k})G^{mJ+j,\,m,\,k} = -\eta_m f_k'\left(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,m,\,k}\right)G^{mJ+j,\,m,\,k}, \tag{16}$$

$$\Delta_k^m\mathbf{v}_i^{mJ+j} = \eta_m\left(O^k - y^{mJ+j,\,m,\,k}\right)f'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,m,\,k})\cdot u_i^{mJ+j} \times g'(\mathbf{v}_i^{mJ+j}\cdot\mathbf{x}^{m,\,k})\mathbf{x}^{m,\,k} = -\eta_m f_k'\left(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,m,\,k}\right) \times u_i^{mJ+j}g'(\mathbf{v}_i^{mJ+j}\cdot\mathbf{x}^{m,\,k})\mathbf{x}^{m,\,k}, \tag{17}$$

$$G^{mJ+j,\,m,\,k} = G(\mathbf{V}^{mJ+j}\mathbf{x}^{m,\,k}), \qquad y^{mJ+j,\,m,\,k} = f(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,m,\,k}), \quad m\in\mathbb{N};\ i = 1, 2, \ldots, n;\ j,\ k = 1, 2, \ldots, J. \tag{18}$$

We mention that OGM-F and OGM-SS are also called cycle learning and almost-cycle learning in Heskes and Wiegerinck (1996), respectively.

3. Main results

For any $\mathbf{x}\in\mathbb{R}^n$, we write $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$, where $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^n$. Let $\Omega_0 = \{\mathbf{w}\in\Omega :\ E_{\mathbf{w}}(\mathbf{w}) = 0\}$ be the stationary point set of the error function $E(\mathbf{w})$, where $\Omega\subset\mathbb{R}^{n(p+1)}$ is a bounded region satisfying (A3) below. Let $\Omega_{0,s}\subset\mathbb{R}$ be the projection of Ω_0 onto the sth coordinate axis, that is,

$$\Omega_{0,s} = \left\{w_s\in\mathbb{R} : \mathbf{w} = (w_1, \ldots, w_s, \ldots, w_{n(p+1)})^T\in\Omega_0\right\} \tag{19}$$

for $s = 1, 2, \ldots, n(p+1)$. To analyze the convergence of the algorithm, we need the following assumptions.

(A1) $g'(t)$ and $f'(t)$ are Lipschitz continuous on any bounded closed interval;
(A2) $\eta_m > 0$, $\sum_{m=0}^{\infty}\eta_m = \infty$, $\sum_{m=0}^{\infty}\eta_m^2 < \infty$;
(A3) There exists a bounded open set $\Omega\subset\mathbb{R}^n$ such that $\{\mathbf{w}^m\}\subset\Omega$ $(m\in\mathbb{N})$;
(A3′) There exists a bounded open set $\Omega'\subset\mathbb{R}^n$ such that $\{\mathbf{u}^m\}\subset\Omega'$ $(m\in\mathbb{N})$, and the derivative of the activation function g in (1) is uniformly bounded and Lipschitz continuous on $\mathbb{R}$.
(A4) $\Omega_{0,s}$ does not contain any interior point for every $s = 1, 2, \ldots, n(p+1)$.

Theorem 3.1. *Assume that conditions* (A1)–(A3) *are valid. Then, starting from an arbitrary initial value* $\mathbf{w}^0$, *the learning sequence* $\{\mathbf{w}^m\}$ *defined by* (9) *and* (10) *or by* (14) *and* (15) *satisfies the following weak convergence*

$$\lim_{m\to\infty}\|E_{\mathbf{w}}\left(\mathbf{w}^m\right)\| = 0. \tag{20}$$

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

$$\lim_{m\to\infty}\mathbf{w}^m = \mathbf{w}^*. \tag{21}$$

Let us make three remarks on the convergence result: (1) We claim that the weak convergence remains valid if the activation function g of the hidden layer is a commonly used sigmoid function and assumptions (A3′) (instead of (A3)) and (A2) are valid. This is due to the fact that the sigmoid function g is uniformly bounded on $\mathbb{R}$ and that (37) is valid even if the weight vectors $\mathbf{v}_i$ $(i = 1, 2, \ldots, n)$ 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 (cf. Zhang et al., 2009). (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 Wu et al. (2005) and Xu et al. (2009), respectively, are special cases of assumption (A4). This relaxed condition makes it much easier to verify the strong convergence in practice.

4. Proofs

For convenience of presentation, we present in detail the convergence proof for OGM-F in the following Section 4.1. Then, in Section 4.2, we briefly point out how to extend the result to OGM-SS.

4.1. Convergence analysis for OGM-F

We first present four useful lemmas for the convergence analysis.

Lemma 4.1. *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*

$$\left|q''(x)\right| \le K, \quad a.e.\ [a, b]. \tag{22}$$

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

$$q(x) \le q(x_0) + q'(x_0)(x - x_0) + C(x - x_0)^2, \quad \forall x_0, x \in [a, b]. \tag{23}$$

Proof. Since $q'(x)$ is Lipschitz continuous on $[a, b]$, $q'(x)$ is absolutely continuous and the derivative $q''(x)$ exists almost everywhere on $[a, b]$. Hence, for almost every $x \in [a, b]$,

$$\begin{aligned}\left|q''(x)\right| &= \left|\lim_{h\to 0}\frac{q'(x+h) - q'(x)}{h}\right| \\ &= \lim_{h\to 0}\left|\frac{q'(x+h) - q'(x)}{h}\right| \le K.\end{aligned} \tag{24}$$

Using the integral Taylor expansion, we deduce that

$$\begin{aligned}q(x) &= q(x_0) + q'(x_0)(x - x_0) \\ &\quad + (x - x_0)^2\int_0^1 (1 - t)q''(x_0 + t(x - x_0))\mathrm{d}t \\ &\le q(x_0) + q'(x_0)(x - x_0) + (x - x_0)^2\int_0^1 K(1 - t)\mathrm{d}t \\ &= q(x_0) + q'(x_0)(x - x_0) + C(x - x_0)^2, \\ &\quad \left(C = \frac{K}{2},\ x_0, x \in [a, b]\right). \quad \square\end{aligned} \tag{25}$$

Lemma 4.2. *Suppose that the learning rate η_m satisfies* (A2) *and that the sequence $\{a_m\}$ $(m \in \mathbb{N})$ satisfies $a_m \ge 0$, $\sum_{m=0}^{\infty}\eta_m a_m^\beta < \infty$ and $|a_{m+1} - a_m| \le \mu\eta_m$ for some positive constants β and μ. Then we have*

$$\lim_{m\to\infty} a_m = 0. \tag{26}$$

Proof. According to (A2), we know that $\eta_m \to 0$ as $m \to \infty$. We claim that $\lim_{k\to\infty}\inf_{m>k} a_m = 0$. Otherwise, if $a_* \equiv \lim_{k\to\infty}\inf_{m>k} a_m \in (0, \infty]$, then by the definition of the inferior limit, there exists an integer $M > k$ such that $a_m \ge \frac{a_*}{2} > 0$ for $m \ge M$, which leads to

$$\sum_{m=0}^{\infty}\eta_m a_m^\beta \ge \left(\frac{a_*}{2}\right)^\beta\sum_{m\ge M}^{\infty}\eta_m = \infty. \tag{27}$$

This contradicts $\sum_{m=0}^{\infty}\eta_m a_m^\beta < \infty$ and confirms the claim. Next, we claim that $\lim_{k\to\infty}\sup_{m>k} a_m = 0$. Otherwise, there exists $\delta \in (0, \infty]$ such that $\lim_{k\to\infty}\sup_{m>k} a_m = \delta$. Then, for any $0 < \varepsilon < \delta$, we can choose two subsequences $\{a_{i_k}\}$ and $\{a_{j_k}\}$ of $\{a_m\}$ to satisfy (1) $a_{i_k} \in \left(0, \frac{\varepsilon}{4}\right)$, $a_{j_k} \in (\varepsilon, \delta)$; (2) $i_k + 1 < j_k < i_{k+1}$; (3) $a_{i_k+1} \in [\frac{\varepsilon}{4}, \frac{\varepsilon}{2}]$. (This can be done because $\lim_{k\to\infty}\inf_{m>k} a_m = 0$, $\lim_{k\to\infty}\sup_{m>k} a_m = \delta$, and $|a_m - a_{m+1}| \le \mu\eta_m \to 0$ as $m \to \infty$.) For any $i_k < m < j_k$, we have $a_m \in [\frac{\varepsilon}{4}, \varepsilon]$. Thus, we conclude that

$$\begin{aligned}\frac{\varepsilon}{2} &\le \left|a_{j_k} - a_{i_k+1}\right| \le |a_{j_k} - a_{j_k-1}| + \cdots + |a_{i_k+2} - a_{i_k+1}| \\ &\le \mu\sum_{m=i_k+1}^{j_k-1}\eta_m \le \mu\sum_{m=i_k+1}^{j_k}\eta_m.\end{aligned}$$

Therefore, we have for all large enough integer k that

$$\sum_{m=i_k}^{j_k}\eta_m a_m^\beta \ge \sum_{m=i_k+1}^{j_k}\eta_m a_m^\beta \ge \left(\frac{\varepsilon}{4}\right)^\beta\sum_{m=i_k+1}^{j_k}\eta_m \ge \frac{2}{\mu}\left(\frac{\varepsilon}{4}\right)^{\beta+1}.$$

But this contradicts $\sum_{m=0}^{\infty}\eta_m a_m^\beta < \infty$ and implies our second claim. Finally, the above two claims together clearly lead to the desired estimate (26). $\square$

Lemma 4.3. *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*

$$S = [\gamma_1, \gamma_2]. \tag{28}$$

Proof. It is obvious that $\gamma_1 \le \gamma_2$ and $S \subseteq [\gamma_1, \gamma_2]$. If $\gamma_1 = \gamma_2$, then (28) 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 $a \in (\gamma_1, \gamma_2)$, 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. $\square$

Let the sequence $\{\mathbf{w}^{mJ+j}\}$ $(m \in \mathbb{N},\ j = 1, 2, \ldots, J)$ be generated by (9) and (10). We introduce the following notations:

$$R^{m,j} = \Delta_j\mathbf{u}^{mJ+j} - \Delta_j\mathbf{u}^{mJ}, \tag{29}$$

$$r_i^{m,j} = \Delta_j\mathbf{v}_i^{mJ+j} - \Delta_j\mathbf{v}_i^{mJ}, \tag{30}$$

$$d^{m,l} = \mathbf{u}^{mJ+l} - \mathbf{u}^{mJ} = \sum_{j=1}^{l}\Delta_j\mathbf{u}^{mJ+j} = \sum_{j=1}^{l}\Delta_j\mathbf{u}^{mJ} + \sum_{j=1}^{l}R^{m,j}, \tag{31}$$

$$h_i^{m,l} = \mathbf{v}_i^{mJ+l} - \mathbf{v}_i^{mJ} = \sum_{j=1}^{l}\Delta_j\mathbf{v}_i^{mJ+j} = \sum_{j=1}^{l}\Delta_j\mathbf{v}_i^{mJ} + \sum_{j=1}^{l}r_i^{m,j}, \tag{32}$$

$$\psi^{m,l,j} = G^{mJ+l,j} - G^{mJ,j}, \tag{33}$$

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

Then, (9) and (10) can be rewritten as

$$\mathbf{u}^{mJ+j} = \mathbf{u}^{mJ} + \sum_{k=1}^{j}\left(\Delta_k\mathbf{u}^{mJ} + R^{m,k}\right), \tag{34}$$

$$\mathbf{v}_i^{mJ+j} = \mathbf{v}_i^{mJ} + \sum_{k=1}^{j}(\Delta_k\mathbf{v}_i^{mJ} + r_i^{m,k}). \tag{35}$$

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

$$\max_{1\le j\le J}\left\{\|\mathbf{x}^j\|,\ |O^j|\right\} = C_1,\ \sup_{m\in\mathbb{N}}\|\mathbf{w}^m\| = C_2. \tag{36}$$

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

Lemma 4.4. *Let conditions* (A1) *and* (A3) *be valid, and let the sequence $\{\mathbf{w}^{mJ+j}\}$ be generated by* (9) *and* (10)*. Then there are constants C_3 –C_7 such that*

$$\left\|G^{mJ+j,k}\right\| \le C_3, \tag{37}$$

$$\|d^{m,l}\| \le C_4\eta_m, \qquad \|\psi^{m,l,j}\| \le C_5\eta_m, \tag{38}$$

$$\|R^{m,j}\| \le C_6\eta_m^2, \qquad \|r_i^{m,j}\| \le C_7\eta_m^2, \tag{39}$$

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

Proof. According to (36), we have

$$|\mathbf{v}_i^{mJ+j}\cdot\mathbf{x}^k|\le\|\mathbf{v}_i^{mJ+j}\|\,\|\mathbf{x}^k\|\le C_1C_2\equiv D_1. \tag{40}$$

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

$$\max_{|t|\le D_1}|g(t)|=C_{3,1}, \tag{41}$$

$$\left\|G^{mJ+j,\,k}\right\|=\left\|G\left(\mathbf{V}^{mJ+j}\mathbf{x}^k\right)\right\|\le\sqrt{n}C_{3,1}\equiv C_3. \tag{42}$$

It follows from (36) and (42) that

$$\left|\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,k}\right|\le\left\|\mathbf{u}^{mJ+j}\right\|\left\|G^{mJ+j,\,k}\right\|\le C_2C_3\equiv D_2. \tag{43}$$

Then, there is a positive constant $C_{4,1}$ such that

$$\max_{|t|\le D_2}\left|f_j'(t)\right|\le C_{4,1}. \tag{44}$$

Furthermore, a combination of (A1), (11), (37) and (40) gives

$$\left\|d^{m,\,l}\right\|=\left\|\mathbf{u}^{mJ+l}-\mathbf{u}^{mJ}\right\|=\left\|\sum_{j=1}^{l}\Delta_j\mathbf{u}^{mJ+j}\right\|\le C_4\eta_m, \tag{45}$$

where $C_4=JC_{4,1}C_3$.

Employing (40), we find that

$$\max_{|t|\le D_1}\left|g'(t)\right|=C_{5,1}, \tag{46}$$

where $C_{5,1}$ is a positive constant. Moreover, we observe that

$$\begin{aligned}\left\|\psi^{m,\,l,\,j}\right\|&=\left\|G^{mJ+l,\,j}-G^{mJ,\,j}\right\|\le\max_{1\le i\le n}\left|g'(t_i)\right|\left\|\mathbf{x}^j\right\|\sum_{i=1}^{n}\|h_i^{m,\,l}\|\\&\le\max_{1\le i\le n}\left|g'(t_i)\right|\left\|\mathbf{x}^j\right\|\sum_{i=1}^{n}\sum_{k=1}^{l}\|\Delta_k\mathbf{v}_i^{mJ+k}\|\\&\le C_5\eta_m,\end{aligned} \tag{47}$$

where $C_5=nlC_{4,1}C_{5,1}\max_{1\le i\le n}|g'(t_i)|\|\mathbf{x}^j\|\sup_{m\in\mathbb{N}}\|\mathbf{w}^n\|\max_{1\le k\le J}\|\mathbf{x}^k\|$, in which $t_i=\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j+\theta_i(\mathbf{v}_i^{mJ+l}-\mathbf{v}_i^{mJ})\cdot\mathbf{x}^j$, $\theta_i\in(0,1)$, and $|t_i|\le\left|\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j\right|+|(\mathbf{v}_i^{mJ+l}-\mathbf{v}_i^{mJ})\cdot\mathbf{x}^j|\le 3C_1C_2$. By virtue of (A1), we see that $\left|g'(t_i)\right|$ $(i=1,2,\ldots,n)$ is bounded.

Combining $f_j'(t)$'s Lipschitz continuity, (36) and (37), we have

$$\begin{aligned}&\left|f_j'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,j})-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ+j,\,j})\right|\\&\quad\le L\left|\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,j}-\mathbf{u}^{mJ}\cdot G^{mJ+j,\,j}\right|\\&\quad\le L\left\|d^{m,\,j}\right\|\left\|G^{mJ+j,\,j}\right\|\le LC_3\left\|d^{m,\,j}\right\|,\end{aligned} \tag{48}$$

$$\begin{aligned}&\left|f_j'(\mathbf{u}^{mJ}\cdot G^{mJ+j,\,j})-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})\right|\\&\quad\le L\left|\mathbf{u}^{mJ}\cdot G^{mJ+j,\,j}-\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right|\\&\quad\le L\left\|\mathbf{u}^{mJ}\right\|\left\|\psi^{m,\,j,\,j}\right\|\le LC_2\left\|\psi^{m,\,j,\,j}\right\|,\end{aligned} \tag{49}$$

where $L>0$ is the Lipschitz constant.

By the definition of $R^{m,\,j}$, we see that

$$\begin{aligned}R^{m,\,j}&=\Delta_j\mathbf{u}^{mJ+j}-\Delta_j\mathbf{u}^{mJ}\\&=-\eta_m\left(f_j'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,j})G^{mJ+j,\,j}-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})G^{mJ,\,j}\right)\\&=-\eta_m\big[f_j'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,j})\psi^{m,\,j,\,j}\\&\quad+\left(f_j'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,\,j})-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ+j,\,j})\right)G^{mJ,\,j}\\&\quad+\left(f_j'(\mathbf{u}^{mJ}\cdot G^{mJ+j,\,j})-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})\right)G^{mJ,\,j}\big].\end{aligned} \tag{50}$$

Therefore, it follows from (37), (38), (48) and (49) that

$$\left\|R^{m,\,j}\right\|\le\eta_m\left(LC_3^2\left\|d^{m,\,j}\right\|+(C_{4,1}+LC_2C_3)\left\|\psi^{m,\,j,\,j}\right\|\right)\le C_6\eta_m^2, \tag{51}$$

where $C_6=\max\left\{LC_3^2C_4,\ (C_{4,1}+LD_2)C_5\right\}$.

Similarly, we can show the existence of a constant $C_7>0$ such that

$$\left\|r_i^{m,\,j}\right\|\le C_7\eta_m^2.\quad\square \tag{52}$$

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

Lemma 4.5. *Let the sequence $\left\{\mathbf{w}^{mJ+j}\right\}$ be generated by (9) and (10). Under assumptions* (A1) *and* (A3)*, there holds*

$$E\left(\mathbf{w}^{(m+1)J}\right)\le E\left(\mathbf{w}^{mJ}\right)-\eta_m\left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2+C_8\eta_m^2,\quad(m=0,1,\ldots) \tag{53}$$

where $C_8>0$ is a constant independent of m and η_m.

Proof. By virtue of assumption (A1) and Lemma 4.1, we know that the derivative $g''(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j+t(h_i^{mJ}\cdot\mathbf{x}^j))$ is integrable almost everywhere on $[0,1]$ and

$$\begin{aligned}&f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\mathbf{u}^{mJ}\cdot\psi^{m,\,J,\,j}\\&=f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\sum_{i=1}^{n}u_i^{mJ}g'(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j)h_i^{m,\,J}\cdot\mathbf{x}^j\\&\quad+f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\sum_{i=1}^{n}u_i^{mJ}(h_i^{m,\,J}\cdot\mathbf{x}^j)^2\cdot\int_0^1(1-t)\\&\quad\times g''(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j+t(h_i^{m,\,J}\cdot\mathbf{x}^j))\mathrm{d}t.\end{aligned} \tag{54}$$

By virtue of Lemma 4.1, (11), (12) and (54), there is a constant $C_9>0$ such that

$$\begin{aligned}&f_j\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j}\right)\\&\le f_j\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\\&\quad+f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j}-\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\\&\quad+C_9\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j}-\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2\\&=f_j(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})\\&\quad+f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})\left(d^{m,\,J}\cdot G^{mJ,\,j}+\mathbf{u}^{mJ}\cdot\psi^{m,\,J,\,j}+d^{m,\,J}\cdot\psi^{m,\,J,\,j}\right)\\&\quad+C_9\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j}-\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2\\&=f_j(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})-\frac{1}{\eta_m}\Delta_j\mathbf{u}^{mJ}\cdot d^{m,\,J}-\frac{1}{\eta_m}\sum_{i=1}^{n}(\Delta_j\mathbf{v}_i^{mJ}\cdot h_i^{m,\,J})\\&\quad+f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\sum_{i=1}^{n}u_i^{mJ}(h_i^{m,\,J}\cdot\mathbf{x}^j)^2\cdot\int_0^1(1-t)\\&\quad\times g''(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j+t(h_i^{m,\,J}\cdot\mathbf{x}^j))\mathrm{d}t\\&\quad+f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})d^{m,\,J}\cdot\psi^{m,\,J,\,j}\\&\quad+C_9\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j}-\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2.\end{aligned} \tag{55}$$

Summing (55) from $j=1$ to $j=J$ and noting (3)–(5), (31) and (32), we have

$$\begin{aligned}&E\left(\mathbf{w}^{(m+1)J}\right)\\&\le E\left(\mathbf{w}^{mJ}\right)-\frac{1}{\eta_m}\left(\left\|\sum_{j=1}^{J}\Delta_j\mathbf{u}^{mJ}\right\|^2+\sum_{i=1}^{n}\left\|\sum_{j=1}^{J}\Delta_j\mathbf{v}_i^{mJ}\right\|^2\right)+\delta_m\\&=E\left(\mathbf{w}^{mJ}\right)-\eta_m\left(\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|^2+\sum_{i=1}^{n}\left\|E_{\mathbf{v}_i}\left(\mathbf{w}^{mJ}\right)\right\|^2\right)+\delta_m\\&=E\left(\mathbf{w}^{mJ}\right)-\eta_m\left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2+\delta_m,\end{aligned} \tag{56}$$

where

$$\begin{aligned}\delta_m = &-\frac{1}{\eta_m}\sum_{j=1}^{J}\Delta_j\mathbf{u}^{mJ}\cdot\sum_{j=1}^{J}R^{m,j}-\frac{1}{\eta_m}\sum_{i=1}^{n}\left(\sum_{j=1}^{J}\Delta_j\mathbf{v}_i^{mJ}\cdot\sum_{j=1}^{J}r_i^{m,j}\right)\\&+\sum_{j=1}^{J}\sum_{i=1}^{n}u_i^{mJ}f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,j}\right)(h_i^{m,J}\cdot\mathbf{x}^j)^2\cdot\int_0^1(1-t)\\&\times g''(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j+t(h_i^{m,J}\cdot\mathbf{x}^j))\mathrm{d}t\\&+\sum_{j=1}^{J}f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,j})d^{m,J}\cdot\psi^{m,J,j}\\&+C_9\sum_{j=1}^{J}\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,j}-\mathbf{u}^{mJ}\cdot G^{mJ,j}\right)^2.\end{aligned}$$

It now follows from (36) and (37) that

$$\begin{aligned}&\left\|G^{mJ,j}\right\|=\left\|G\left(\mathbf{V}^{mJ}\mathbf{x}^j\right)\right\|\le C_3,\\&\left|\mathbf{u}^{mJ}\cdot G^{mJ,j}\right|\le\left\|\mathbf{u}^{mJ}\right\|\left\|G^{mJ,j}\right\|\le C_2C_3=D_2.\end{aligned}\tag{57}$$

By (11), (42)–(44) and (51), the first term of δ_m can be estimated as follows.

$$\begin{aligned}&\left\|-\frac{1}{\eta_m}\sum_{j=1}^{J}\Delta_j\mathbf{u}^{mJ}\cdot\sum_{j=1}^{J}R^{m,j}\right\|\\&\le\frac{1}{\eta_m}\sum_{j=1}^{J}\left\|\Delta_j\mathbf{u}^{mJ}\right\|\cdot\sum_{j=1}^{J}\left\|R^{m,j}\right\|\le C_{8,1}\eta_m^2,\end{aligned}\tag{58}$$

where $C_{8,1}=J^2C_3C_{4,1}C_6=JC_4C_6$.

Similar estimates for the other terms of δ_m can be obtained with corresponding constants $C_{8,t}>0$ for $t=2,\ldots,5$. Finally, the desired estimate (53) is proved by setting $C_8=\sum_{t=1}^{5}C_{8,t}$. □

Now, we are ready to prove the convergence theorem.

Proof of Theorem 3.1 for OGM-F. The proof is divided into two parts, dealing with (20) and (21), respectively.

Proof of (20). By (A2) and Lemma 4.5, we conclude that

$$\begin{aligned}&\sum_{m=0}^{\infty}\eta_m\left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2\\&=\sum_{m=0}^{\infty}\eta_m\left(\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|^2+\left\|E_{\mathbf{v}}\left(\mathbf{w}^{mJ}\right)\right\|^2\right)<\infty,\end{aligned}\tag{59}$$

$$\sum_{m=0}^{\infty}\eta_m\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|^2<\infty.\tag{60}$$

Employing the integral Taylor expansion, we deduce that

$$\begin{aligned}&f_j'(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,j})G^{(m+1)J,j}-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,j})G^{mJ,j}\\&=f_j'(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,j})\psi^{m,J,j}\\&\quad+\left(f_j'(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,j})-f_j'(\mathbf{u}^{mJ}\cdot G^{(m+1)J,j})\right)G^{mJ,j}\\&\quad+\left(f_j'(\mathbf{u}^{mJ}\cdot G^{(m+1)J,j})-f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,j})\right)G^{mJ,j}\\&=f_j'(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,j})\psi^{m,J,j}\\&\quad+(d^{m,J}\cdot G^{(m+1)J,j})G^{mJ,j}\cdot\int_0^1(1-t)\\&\quad\times f_j''(\mathbf{u}^{mJ}\cdot G^{(m+1)J,j}+t(d^{m,J}\cdot G^{(m+1)J,j}))\mathrm{d}t\\&\quad+(\mathbf{u}^{mJ}\cdot\psi^{m,J,j})G^{mJ,j}\cdot\int_0^1(1-t)\\&\quad\times f_j''(\mathbf{u}^{mJ}\cdot G^{mJ,j}+t(\mathbf{u}^{mJ}\cdot\psi^{m,J,j}))\mathrm{d}t.\end{aligned}\tag{61}$$

Note (A2) and let $\eta_c>0$ be an upper bound of $\{\eta_m\}_{m=0}^{\infty}$. It follows from (36)–(38) that

$$\begin{aligned}&\left|\mathbf{u}^{mJ}\cdot G^{(m+1)J,j}+t\left(d^{m,J}\cdot G^{(m+1)J,j}\right)\right|\\&\le\left(\left\|\mathbf{u}^{mJ}\right\|+\left\|d^{m,J}\right\|\right)\left\|G^{(m+1)J,j}\right\|\\&\le(C_2+C_4\eta_c)\,C_3,\quad t\in(0,1),\end{aligned}\tag{62}$$

$$\begin{aligned}&\left|\mathbf{u}^{mJ}\cdot G^{mJ,j}+t\left(\mathbf{u}^{mJ}\cdot\psi^{m,J,j}\right)\right|\\&\le\left\|\mathbf{u}^{mJ}\right\|\left(\left\|G^{mJ,j}\right\|+\left\|\psi^{m,J,j}\right\|\right)\\&\le C_2(C_3+C_5\eta_c)=D_2+C_2C_5\eta_c,\quad t\in(0,1).\end{aligned}\tag{63}$$

According to (62), (63) and the proof of Lemma 4.1, there are positive constants C_{10}, $C_{11}>0$ such that

$$\left|\int_0^1(1-t)f_j''\left(\mathbf{u}^{mJ}\cdot G^{(m+1)J,j}+td^{m,J}\cdot G^{(m+1)J,j}\right)\mathrm{d}t\right|\le C_{10},\tag{64}$$

$$\left|\int_0^1(1-t)f_j''\left(\mathbf{u}^{mJ}\cdot G^{mJ,j}+t\mathbf{u}^{mJ}\cdot\psi^{m,J,j}\right)\mathrm{d}t\right|\le C_{11}.\tag{65}$$

By (43), we obtain $\left|\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,j}\right|\le C_2C_3=D_2$. Employing (4) and (37), (38), (44), (64) and (65), and summing (61) from $j=1$ to $j=J$, we conclude that

$$\begin{aligned}&\left|\left\|E_{\mathbf{u}}\left(\mathbf{w}^{(m+1)J}\right)\right\|-\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|\right|\le\left\|E_{\mathbf{u}}\left(\mathbf{w}^{(m+1)J}\right)-E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|\\&\le C_{10}J\max_{\substack{1\le j\le J\\m\in\mathbb{N}}}\left(\left\|G^{(m+1)J,j}\right\|\left\|G^{mJ,j}\right\|\right)\left\|d^{m,J}\right\|\\&\quad+\left(JC_{4,1}+C_{11}J\max_{\substack{1\le j\le J\\m\in\mathbb{N}}}\left(\left\|\mathbf{u}^{mJ}\right\|\left\|G^{mJ,j}\right\|\right)\right)\max_{\substack{1\le j\le J\\m\in\mathbb{N}}}\left\|\psi^{m,J,j}\right\|\\&\le C_{12}\eta_m,\end{aligned}\tag{66}$$

where $C_{12}=JC_3^2C_4C_{10}+JC_{4,1}C_5+JD_2C_5C_{11}$. Combining (59), (66) and Lemma 4.2 results in $\lim_{m\to\infty}\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|=0$.

Similarly as in the proof to (66), there exists a positive constant C_{13} such that

$$\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ+j}\right)-E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|\le C_{13}\eta_m.\tag{67}$$

Since

$$\begin{aligned}\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ+j}\right)\right\|&\le\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ+j}\right)-E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|+\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|\\&\le C_{13}\eta_m+\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|,\end{aligned}\tag{68}$$

we have $\lim_{m\to\infty}\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ+j}\right)\right\|=0$ for $j=1,2,\ldots,J$. Similarly, we deduce that $\lim_{m\to\infty}\left\|E_{\mathbf{v}_i}\left(\mathbf{w}^{mJ+j}\right)\right\|=0$ for $i=1,\ldots,n$, $j=1,2,\ldots,J$, and

$$\lim_{m\to\infty}\left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ+j}\right)\right\|=0,\quad j=1,2,\ldots,J.\tag{69}$$

This immediately gives

$$\lim_{m\to\infty}\|E_{\mathbf{w}}\left(\mathbf{w}^{m}\right)\|=0.\quad\square\tag{70}$$

Proof of (21). According to (A3), the sequence $\{\mathbf{w}^m\}$ ($m\in\mathbb{N}$) has a subsequence $\{\mathbf{w}^{m_k}\}$ ($k\in\mathbb{N}$) that is convergent to, say, $\mathbf{w}^*\in\Omega_0$. It follows from (20) and the continuity of $E_{\mathbf{w}}(\mathbf{w})$ that

$$\left\|E_{\mathbf{w}}\left(\mathbf{w}^*\right)\right\|=\lim_{k\to\infty}\left\|E_{\mathbf{w}}\left(\mathbf{w}^{m_k}\right)\right\|=\lim_{m\to\infty}\left\|E_{\mathbf{w}}\left(\mathbf{w}^{m}\right)\right\|=0.\tag{71}$$

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

Next, by reduction to absurdity, we prove that $\{\mathbf{w}^m\}$ has precisely one accumulation point. Let us assume to the contrary that $\{\mathbf{w}^m\}$ has at least two accumulation points $\overline{\mathbf{w}} \neq \widetilde{\mathbf{w}}$. We write $\mathbf{w}^m = (w_1^m, w_2^m, \ldots, w_{n(p+1)}^m)^T$. It is easy to see from (9)–(12) that $\lim_{m\to\infty} \left\|\mathbf{w}^{m+1} - \mathbf{w}^m\right\| = 0$, or equivalently, $\lim_{m\to\infty} |w_i^{m+1} - w_i^m| = 0$ for $i = 1, 2, \ldots, n(p+1)$. Without loss of generality, we assume that the first components of $\overline{\mathbf{w}}$ and $\widetilde{\mathbf{w}}$ do not equal to each other, that is, $\bar{w}_1 \neq \tilde{w}_1$. For any real number $\lambda \in (0, 1)$, let $w_1^\lambda = \lambda\bar{w}_1 + (1-\lambda)\tilde{w}_1$. By Lemma 4.3, there exists a subsequence $\left\{w_1^{m_{k_1}}\right\}$ of $\left\{w_1^m\right\}$ 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 $\{m_{k_1}\} \supset \{m_{k_2}\} \supset \cdots \supset \{m_{k_{n(p+1)}}\}$ with $w_i^\lambda = \lim_{k_i\to\infty} w_i^{m_{k_i}}$ for each $i = 1, 2, \ldots, n(p+1)$. Write $\mathbf{w}^\lambda = (w_1^\lambda, w_2^\lambda, \ldots, w_{n(p+1)}^\lambda)^T$. Then, we see that $\mathbf{w}^\lambda$ is an accumulation point of $\{\mathbf{w}^m\}$ for any $\lambda \in (0, 1)$. But this means that $\Omega_{0,1}$ has interior points, which contradicts (A4). Thus, $\mathbf{w}^*$ must be a unique accumulation point of $\{\mathbf{w}^m\}_{m=0}^\infty$. This completes the proof of the strong convergence. □

4.2. Convergence analysis for OGM-SS

Now, let the sequence $\left\{\mathbf{w}^{mJ+j}\right\}$ $(m \in \mathbb{N},\ j = 1, 2, \ldots, J)$ be generated by (14) and (15), and let

$$R^{m,j} = \Delta_j^m \mathbf{u}^{mJ+j} - \Delta_j^m \mathbf{u}^{mJ}, \tag{72}$$

$$r_i^{m,j} = \Delta_j^m \mathbf{v}_i^{mJ+j} - \Delta_j^m \mathbf{v}_i^{mJ}, \tag{73}$$

$$d^{m,l} = \mathbf{u}^{mJ+l} - \mathbf{u}^{mJ} = \sum_{j=1}^{l} \Delta_j^m \mathbf{u}^{mJ+j} = \sum_{j=1}^{l} \Delta_j^m \mathbf{u}^{mJ} + \sum_{j=1}^{l} R^{m,j}, \tag{74}$$

$$h_i^{m,l} = \mathbf{v}_i^{mJ+l} - \mathbf{v}_i^{mJ} = \sum_{j=1}^{l} \Delta_j^m \mathbf{v}_i^{mJ+j} = \sum_{j=1}^{l} \Delta_j^m \mathbf{v}_i^{mJ} + \sum_{j=1}^{l} r_i^{m,j}, \tag{75}$$

$$\psi^{m,l,j} = G^{mJ+l,m,j} - G^{mJ,m,j}, \tag{76}$$

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

It is obvious that Lemmas 4.1–4.3 are not influenced by the new definitions. In place of Lemmas 4.4 and 4.5, we now have the following two Lemmas.

Lemma 4.6. *Let conditions* (A1) *and* (A3) *be valid, and let the sequence* $\left\{\mathbf{w}^{mJ+j}\right\}$ *be generated by* (14) *and* (15). *Then, there hold the following estimates with the same constants* C_3 *–*C_7 *as in Lemma* 4.4*:*

$$\left\|G^{mJ+j,m,k}\right\| \leq C_3, \tag{77}$$

$$\|d^{m,l}\| \leq C_4\eta_m, \qquad \|\psi^{m,l,j}\| \leq C_5\eta_m, \tag{78}$$

$$\|R^{m,j}\| \leq C_6\eta_m^2, \qquad \|r_i^{m,j}\| \leq C_7\eta_m^2, \tag{79}$$

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

Proof. According to (36), we have

$$|\mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^{m,k}| \leq \|\mathbf{v}_i^{mJ+j}\| \max_{1\leq k\leq J} \left\|\mathbf{x}^k\right\| \leq C_1C_2 \equiv D_1. \tag{80}$$

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

$$\max_{|t|\leq D_1} |g(t)| = C_{3,1}, \tag{81}$$

$$\left\|G^{mJ+j,m,k}\right\| = \left\|G\left(\mathbf{V}^{mJ+j}\mathbf{x}^{m,k}\right)\right\| \leq \sqrt{n}C_{3,1} \equiv C_3. \tag{82}$$

Similarly, (78) and (79) can be proved after adjusting the corresponding superscripts in the proof to Lemma 4.4. □

Lemma 4.7. *Let the sequence* $\left\{\mathbf{w}^{mJ+j}\right\}$ *be generated by* (14) *and* (15). *Under assumptions* (A1)*and* (A3), *there holds*

$$E\left(\mathbf{w}^{(m+1)J}\right) \leq E\left(\mathbf{w}^{mJ}\right) - \eta_m \left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2 + C_8\eta_m^2, \quad (m = 0, 1, \ldots) \tag{83}$$

where $C_8 > 0$ *is the same constant defined in Lemma* 4.5.

Proof. As in the proof to Lemma 4.6, we only need to adjust some superscripts. For example, corresponding to (54), we change the related superscripts and get

$$\begin{aligned} & f_j'\left(\mathbf{u}^{mJ} \cdot G^{mJ,m,j}\right)\mathbf{u}^{mJ} \cdot \psi^{m,J,j} \\ &= f_j'\left(\mathbf{u}^{mJ} \cdot G^{mJ,m,j}\right)\sum_{i=1}^{n} u_i^{mJ} g'(\mathbf{v}_i^{mJ} \cdot \mathbf{x}^{m,j}) h_i^{m,J} \cdot \mathbf{x}^{m,j} \\ &\quad + f_j'\left(\mathbf{u}^{mJ} \cdot G^{mJ,m,j}\right)\sum_{i=1}^{n} u_i^{mJ}(h_i^{m,J} \cdot \mathbf{x}^{m,j})^2 \\ &\quad \times \int_0^1 (1-t) g''(\mathbf{v}_i^{mJ} \cdot \mathbf{x}^{m,j} + t(h_i^{m,J} \cdot \mathbf{x}^{m,j}))\mathrm{d}t. \end{aligned} \tag{84}$$

The details are left to the interested readers. □

Proof of Theorem 3.1 for OGM-SS. We can use Lemmas 4.1–4.3 and Lemmas 4.6–4.7 to obtain the weak and strong convergence results for OGM-SS precisely as in the proof to Theorem 3.1 for OGM-F. □

5. Conclusions

In this paper, we present a comprehensive study on the weak and strong convergence for three-layer BP neural networks. Compared with the existing convergence results, the corresponding assumptions are more relaxed. Our convergence analysis holds for more extensive BP neural networks, e.g., S–S, S–P, P–P and P–S type neural networks.

References

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Chakraborty, D., & Pal, N. R. (2003). A novel training scheme for multilayered perceptrons to realize proper generalization and incremental learning. *IEEE Transactions on Neural Networks*, *14*, 1–14.

Fine, T. L., & Mukherjee, S. (1999). Parameter convergence and learning curves for neural networks. *Neural Computation*, *11*, 747–769.

Finnoff, W. (1994). Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima. *Neural Computation*, *6*, 285–295.

Heskes, T., & Wiegerinck, W. (1996). A theoretical comparison of batch-mode, online, cyclic, and almost-cyclic learning. *IEEE Transactions on Neural Networks*, *7*, 919–925.

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

Li, Z. X., & Ding, X. S. (2005). Prediction of stock market by bp neural networks with technical indexes as input. *Numerical Mathematics: A Journal of Chinese Universities*, *27*, 373–377.

Li, Z. X., Wu, W., & Tian, Y. L. (2004). Convergence of an online gradient method for feedforward neural networks with stochastic inputs. *Journal of Computational and Applied Mathematics*, *163*, 165–176.

Liang, Y. C., Feng, D. P., Lee, H. P., Lim, S. P., & Lee, K. H. (2002). Successive approximation training algorithm for feedforward neural networks. *Neurocomputing*, *42*, 11–322.

Mangasarian, O. L., & Solodov, M. V. (1994). Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. *Optimization Methods and Software*, *4*, 117–134.

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

Parker, D. B. (1982). Learning-logic, invention report. Stanford University, Stanford, Calif.

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

Shao, H. M., Wu, W., & Liu, W. B. (2007). Convergence of BP algorithm for training MLP with linear output. *Numerical Mathematics: A Journal of Chinese Universities (English Series)*, *16*, 193–202.

Tadic, V., & Stankovic, S. (2000). Learning in neural networks by normalized stochastic gradient algorithm: local convergence. In *Proceedings of the 5th seminar neural networks application electronic engineering*.

Terence, D. S. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, *2*, 459–473.

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

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

Wu, W., & Xu, Y. S. (2002). Deterministic convergence of an on-line gradient method for neural networks. *Journal of Computational and Applied Mathematics*, *144*, 335–347.

Wu, W., Feng, G. R., Li, Z. X., & Xu, Y. S. (2005). Deterministic convergence of an online gradient method for BP neural networks. *IEEE Transactions on Neural Networks*, *16*, 533–540.

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

Wu, W., & Shao, Z. Q. (2003). Convergence of online gradient methods for continuous perceptrons with linearly separable training patterns. *Applied Mathematics Letters*, *16*, 999–1002.

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

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

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