Convergence of online gradient method for feedforward neural networks with smoothing $L_{1/2}$ regularization penalty[☆]

Qinwei Fan[a,b], Jacek M. Zurada[b,c], Wei Wu[a,*]

[a] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China*
[b] *Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA*
[c] *Spoleczna Akademia Nauk, 90-011 Lodz, Poland*

Abstract

Minimization of the training regularization term has been recognized as an important objective for sparse modeling and generalization in feedforward neural networks. Most of the studies so far have been focused on the popular L_2 regularization penalty. In this paper, we consider the convergence of online gradient method with smoothing $L_{1/2}$ regularization term. For normal $L_{1/2}$ regularization, the objective function is the sum of a non-convex, non-smooth, and non-Lipschitz function, which causes oscillation of the error function and the norm of gradient. However, using the smoothing approximation techniques, the deficiency of the normal $L_{1/2}$ regularization term can be addressed. This paper shows the strong convergence results for the smoothing $L_{1/2}$ regularization. Furthermore, we prove the boundedness of the weights during the network training. The assumption that weights are bounded is no longer needed for the proof of convergence. Simulation results support the theoretical findings and demonstrate that our algorithm has better performance than two other algorithms with L_2 and normal $L_{1/2}$ regularizations respectively.

Key words: Feedforward neural networks, Online gradient method, Smoothing $L_{1/2}$ regularization, Boundedness, Convergence

[☆]This work was supported by National Science Foundation of China (Numbers 11171367 and 91230103) and China Scholarship Council (CSC).

[*]Corresponding author.
Email address: wuweiw@dlut.edu.cn (Wei Wu[a,*])

1. Introduction

Multilayer feedforward neural networks (FNN) theories and applications have been dominating in neural network literature for the last two decades. Researchers have studied the properties of such networks and their training methods with popular algorithm known as the error backpropagation (BP) [1–4]. Based on the error correction learning rule, the BP algorithm can be viewed as a generalization of the least mean square (LMS) learning law. It is an approximate steepest descent algorithm, in which mean square error is the performance index.

There are two practical ways to implement the gradient method: the batch gradient learning, in which the network weights are updated after all training examples have been processed by the network, and the online gradient learning, in which weights are updated immediately after each training example is processed [5, 6].

The online gradient method (OGM) is a more popular algorithm compared with the batch gradient learning algorithm for BP training [7–9]. Furthermore, OGM is more efficient in terms of both storage and computation time, especially when the task to be learnt is non-stationary and based on instantaneous modifications of network parameters calculated only for the latest training example in a sequence . This process is inherently stochastic because a new training example is selected at random each time after the training error is determined. This is to be contrasted with the batch learning, in which all training examples are aggregated to determine the training error, leading to a deterministic algorithm [10, 11]. There are also other effective methods to train FNN, conjugate gradient method as reported in [12], the extreme learning machine algorithm based on the least-squares for FNN in many applications [13, 14].

Although BP training algorithm has been known for many years, it is well known that the general drawbacks of gradient-based BP training schemes are their more likely divergence and weaker generalization [2]. In the procedure of training the FNN with the sum-squared error (SSE) function, a generally used cost function in neural networks [2, 15, 16], the weights sometimes become very large and overfitting tends to occur. In order to solve this problem, a standard technique is to add an extra regularization term also called the penalty term [17–24].

There are four classical different penalty terms for BPNN: The first one is

weight decay [18], and the error function with the penalty term is defined by

$$E(W) = \tilde{E}(W) + \lambda \sum_{i \in C} w_i^2 = \frac{1}{2} \sum_{j=1}^{J} e_j^2 + \lambda \|W\|^2 \tag{1.1}$$

where $\tilde{E}(W) = \sum_{j=1}^{J} e_j^2$ is the conventional square error function for measuring the accuracy of the network, J is the number of training samples, e_j is the error between the network output value and the target value for the j-th sample, C is the set of all weights, λ is the penalty coefficient, and $\|\cdot\|$ stands for the Euclidean norm. The three other penalty terms for BPNN are weight elimination [25] , approximate smoother [26] and inner-product form [27]. In the pruning procedure, all the penalty (or regularization) term is expressed as the squared norm of the weights [18, 28], but these penalty methods do not work very well for the purpose of driving unnecessary weights to zero. In other words, the network is still not sparse enough in the sense of exact sparse network structure. So how to efficiently prune the network structure becomes our main objective.

It is well known that variable selection and feature extraction are basic problems in high dimensional and massive data analysis, but the traditional variable selection method such as AIC, BIC and CP [29–31] are infeasible for high dimensional data. Thus innovative variable selection procedure has been a hot topic in machine learning, and regularization methods recently have been used as feasible approaches to solve the problem. A class of popular regularization methods has the form:

$$\min\{\frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + \lambda \|f\|_k\} \tag{1.2}$$

where $l(.,.)$ is a loss function, $(x_i, y_i)_{i=1}^n$ is a data set, and λ is the regularization parameter. When f is linear and the loss function is a square loss, $\|f\|_k$ is normally taken as the norm of the coefficient of linear model.

In [32], a proposed $L_{1/2}$ regularizer has been shown to have many promising properties such as unbiasedness, sparsity and oracle properties. Also the experiments have shown that the $L_{1/2}$ regularizer can be very useful and efficient, and can be taken as a representative of the L_r $(0 < r < 1)$ regularizer [33–36]. The $L_{1/2}$ regularizer has the following form:

$$\hat{\beta}_{L_{\frac{1}{2}}} = argmin\{\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{i=1}^{n} |\beta_i|^{\frac{1}{2}}\} \tag{1.3}$$

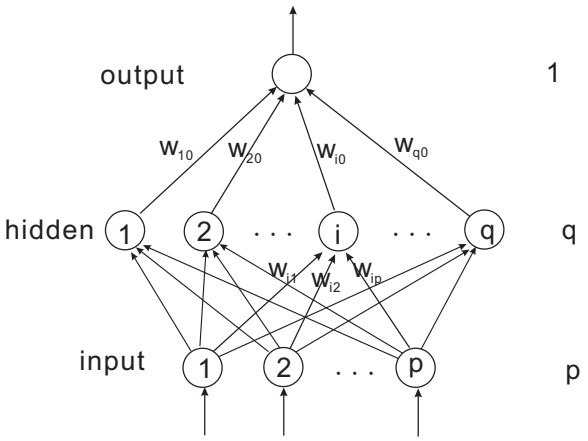

Figure 1: Feedforward neural network with one hidden layer and one output

where λ is the regularization parameter.

In this paper, the $L_{1/2}$ regularization term is introduced for the online gradient learning algorithm for the pruning of FNN. In doing so, we notice that because the usual $L_{1/2}$ regularization term is not smooth at the origin, this leads to difficulty in convergence and, more importantly, oscillation in numerical computation are observed in our numerical experiments. To overcome this drawback, a modified smoothing $L_{1/2}$ regularization term at the origin is introduced. The experiments show that the $L_{1/2}$ regularization algorithm with smoothing is effective, efficient, and performs much better than other algorithms with respect to many aspects. We also prove that the error function is decreasing monotonically, and the online gradient method with $L_{1/2}$ smoothing regularization term is deterministically convergent.

The remaining part of this paper is organized as follows. The online gradient method with L_2 regularization (OGL2), the $L_{1/2}$ regularization penalty term (OGL1/2) and the smoothing $L_{1/2}$ regularization penalty term (OGSL1/2) is described in Section 2. In Section 3, selected convergence results of OGSL1/2 are presented, with their proofs are given in the Appendix. Supporting numerical experiments are presented in Section 4.

2. Algorithm description

Let us begin with an introduction of an FNN with three layers. The neuron numbers of the input, hidden and output layers are p, q and 1, respectively (see Figure.1). Let $\{\xi^j, O^j\}_{j=1}^J \subset \mathbb{R}^p \times \mathbb{R}$ be a given set of training samples, where

ξ^j and O^j are the input and the corresponding ideal output of the j-th sample, respectively. Let $w_i = (w_{i1}, w_{i2}, \cdots, w_{ip}) \in \mathbb{R}^p$ be the weight vector between the input and the hidden layers for $i = 1, 2, \cdots, q$, and let $w_0 = (w_{10}, w_{20}, \cdots, w_{q0}) \in \mathbb{R}^q$ be the weight vector connecting the hidden and the output layers. We write all weights in a compact form, i.e., $W = (w_0, w_1, \cdots, w_q) \in \mathbb{R}^{q+pq}$, and we define a matrix $V = (w_1^T, w_2^T, \cdots, w_q^T)^T \in \mathbb{R}^{q\times p}$. Let $g : \mathbb{R} \to \mathbb{R}$ be a given transfer function for the hidden and output layers. For convenience, we introduce the following vector valued function $G : \mathbb{R}^q \to \mathbb{R}^q$:

$$G(x) = (g(x_1), g(x_2), \cdots, g(x_q))^T, \quad \forall x \in \mathbb{R}^q$$

Then, for any input $\xi \in \mathbb{R}^p$, the actual output vector of the hidden neurons is $G(V\xi)$, and the actual output is

$$\zeta = g(w_0 \cdot G(V\xi)) \tag{2.1}$$

We remark that the bias term has been ignored in the description (2.1) of the algorithm for simplicity. Inclusion of this term will not cause any trouble in the analysis.

2.1. Online gradient method with L_2 regularization (OGL2)

For fixed weights W, a conventional square error function is given by

$$\tilde{E}(W) = \frac{1}{2}\sum_{j=1}^{J}(O^j - g(w_0 \cdot G(V\xi^j)))^2 \tag{2.2}$$

By adding the L_2 regularization penalty term, the total error function takes the following form

$$\begin{aligned} E(W) &= \tilde{E}(W) + \lambda \sum_{i=0}^{q} \|w_i\|^2 \\ &= \sum_{j=1}^{J} g_j(w_0 \cdot G(V\xi^j)) + \lambda \sum_{i=0}^{q} \|w_i\|^2 \end{aligned} \tag{2.3}$$

where $g_j(t) := \frac{1}{2}(O^j - g(t))^2$ is a composite function, the derivative of this function is as follows:

$$g_j'(t) := -(O^j - g(t)) \cdot g'(t).$$

The gradient of the error function with respect to W is given by

$$E_W(W) = (E_{w_0}(W), E_{w_1}(W), \cdots, E_{w_q}(W))^T \tag{2.4}$$

where

$$E_{w_0}(W) = \sum_{j=1}^{J} g_j'(w_0 \cdot G(V\xi^j))G(V\xi^j) + 2\lambda w_0 \tag{2.5}$$

$$E_{w_i}(W) = \sum_{j=1}^{J} g_j'(w_0 \cdot G(V\xi^j))w_{i0}g'(w_i \cdot \xi^j)\xi^j + 2\lambda w_i \tag{2.6}$$

where $i = 1, 2, \cdots, q$.

Given an arbitrary initial weight vector $W^0 \in \mathbb{R}^{q+pq}$, the online gradient method with L_2 regularization updates the weights $\{W^n\}$ after the presentation of each training sample ξ^j. As done in [37], we choose the training samples in a fixed order and the OGL2 can be described as follows:

$$W^{nJ+j} = W^{nJ+j-1} - \eta_n \Delta_j^n W^{nJ+j-1} \tag{2.7}$$

where $n = 0, 1, 2, \cdots$; $j = 1, 2, \cdots, J$; η_n is the learning rate in the n-th training epoch; and

$$\begin{aligned} \Delta_j^n w_0^{nJ+j-1} &= g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))G(V^{nJ+j-1}\xi^j) \\ &\quad + \frac{2\lambda}{J} w_0^{nJ+j-1} \end{aligned} \tag{2.8}$$

$$\begin{aligned} \Delta_j^n w_i^{nJ+j-1} &= g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))w_{i0}^{nJ+j-1} \\ &\quad \times g'(w_i^{nJ+j-1} \cdot \xi^j)\xi^j + \frac{2\lambda}{J} w_i^{nJ+j-1} \end{aligned} \tag{2.9}$$

where $n = 0, 1, 2, \cdots$; $j = 1, 2, \cdots, J$.

2.2. Online gradient method with $L_{1/2}$ regularization (OGL1/2)

The error function with the $L_{1/2}$ regularization penalty term is

$$\begin{aligned} E(W) &= \frac{1}{2}\sum_{j=1}^{J}(O^j - g(w_0 \cdot G(V\xi^j)))^2 + \lambda \sum_{i=1}^{q}\sum_{k=0}^{p} |w_{ik}|^{\frac{1}{2}} \\ &= \sum_{j=1}^{J} g_j(w_0 \cdot G(V\xi^j)) + \lambda \sum_{i=1}^{q}\sum_{k=0}^{p} |w_{ik}|^{\frac{1}{2}} \end{aligned} \tag{2.10}$$

where $g_j(t) := \frac{1}{2}(O^j - g(t))^2$ and $|\cdot|$ denotes the absolute value. The gradient error function depending on W is given by

$$\begin{aligned} E_W(W) =&(E_{w_{10}}(W), E_{w_{20}}(W), \cdots, E_{w_{q0}}(W), \\ &E_{w_{11}}(W), E_{w_{12}}(W), \cdots, E_{w_{1p}}(W), \\ &E_{w_{21}}(W), E_{w_{22}}(W), \cdots, E_{w_{2p}}(W), \\ &\cdots, E_{w_{q1}}(W), E_{w_{q2}}(W), \cdots, E_{w_{qp}}(W))^T \end{aligned} \tag{2.11}$$

where

$$E_{w_{i0}}(W) = \sum_{j=1}^{J} g_j'(w_0 \cdot G(V\xi^j))g(w_i\xi^j) + \frac{\lambda sgn(w_{i0})}{2|w_{i0}|^{\frac{1}{2}}} \tag{2.12}$$

$$E_{w_{ik}}(W) = \sum_{j=1}^{J} g_j'(w_0 \cdot G(V\xi^j))w_{i0}g'(w_i \cdot \xi^j)\xi_k^j + \frac{\lambda sgn(w_{ik})}{2|w_{ik}|^{\frac{1}{2}}} \tag{2.13}$$

where $i = 1, 2, \cdots, q$; and $k = 1, 2, \cdots, p$.

Given an initial weight $W^0 \in \mathbb{R}^{q+pq}$, OGL1/2 updates the weights $\{W^n\}$ iteratively by

$$W^{nJ+j} = W^{nJ+j-1} - \eta_n \Delta_j^n W^{nJ+j-1} \tag{2.14}$$

where $n = 0, 1, 2, \cdots$; $j = 1, 2, \cdots, J$; and

$$\begin{aligned} \Delta_j^n w_{i0}^{nJ+j-1} &= g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))g(w_i^{nJ+j-1}\xi^j) \\ &\quad + \frac{\lambda sgn(w_{i0}^{nJ+j-1})}{2J|w_{i0}^{nJ+j-1}|^{\frac{1}{2}}} \end{aligned} \tag{2.15}$$

$$\begin{aligned} \Delta_j^n w_{ik}^{nJ+j-1} &= g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))w_{i0}^{nJ+j-1} \\ &\quad \times g'(w_i^{nJ+j-1} \cdot \xi^j)\xi_k^j + \frac{\lambda sgn(w_{ik}^{nJ+j-1})}{2J|w_{ik}^{nJ+j-1}|^{\frac{1}{2}}} \end{aligned} \tag{2.16}$$

where $i = 1, 2, \cdots, q$; $k = 1, 2, \cdots, p$; and η_n is the learning rate in the n-th training epoch.

2.3. Online gradient method with smoothing $L_{1/2}$ regularization (OGSL1/2)

To overcome the drawback of error oscillation and difficulties in convergence analysis, a modified $L_{1/2}$ regularization term is proposed to smooth the usual one at the origin. This results in the following error function with a smoothing $L_{1/2}$ regularization penalty term:

$$\begin{aligned} E(W) &= \frac{1}{2}\sum_{j=1}^{J}(O^j - g(w_0 \cdot G(V\xi^j)))^2 + \lambda \sum_{i=1}^{q}\sum_{k=0}^{p} f(w_{ik})^{\frac{1}{2}} \\ &= \sum_{j=1}^{J} g_j(w_0 \cdot G(V\xi^j)) + \lambda \sum_{i=1}^{q}\sum_{k=0}^{p} f(w_{ik})^{\frac{1}{2}} \end{aligned} \tag{2.17}$$

In order to approximate the non-smooth function $|x|$, we use the smoothing function $f(x)$ defined by

$$f(x) = \begin{cases} |x|, & if\ |x| \geq a \\ -\frac{1}{8a^3}x^4 + \frac{3}{4a}x^2 + \frac{3}{8}a, & if\ |x| < a \end{cases} \tag{2.18}$$

where a is a small positive constant. Then, we can easily show that

$$f(x) \in [\frac{3}{8}a, +\infty), \quad f'(x) \in [-1, 1], \quad f''(x) \in [0, \frac{3}{2a}]$$

The gradient of the error function (2.17) with respect to W can be written in the form of (2.11) with

$$E_{w_{i0}}(W) = \sum_{j=1}^{J} g'_j(w_0 \cdot G(V\xi^j))g(w_i \cdot \xi^j) + \lambda \frac{f'(w_{i0})}{2f(w_{i0})^{\frac{1}{2}}} \tag{2.19}$$

$$E_{w_{ik}}(W) = \sum_{j=1}^{J} g'_j(w_0 \cdot G(V\xi^j))w_{i0}g'(w_i \cdot \xi^j)\xi_k^j + \lambda \frac{f'(w_{ik})}{2f(w_{ik})^{\frac{1}{2}}} \tag{2.20}$$

where $i = 1, 2, \cdots, q$; and $k = 1, 2, \cdots, p$.

Given an initial weight $W^0 \in \mathbb{R}^{q+pq}$, $OGSL1/2$ updates the weights $\{W^n\}$ iteratively by

$$W^{nJ+j} = W^{nJ+j-1} - \eta_n \Delta_j^n W^{nJ+j-1} \tag{2.21}$$

where $n = 0, 1, 2, \cdots$; $j = 1, 2, \cdots, J$; and

$$\Delta_j^n w_{i0}^{nJ+j-1} = g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))g(w_i^{nJ+j-1} \cdot \xi^j) + \lambda \frac{f'(w_{i0}^{nJ+j-1})}{2Jf(w_{i0}^{nJ+j-1})^{\frac{1}{2}}} \tag{2.22}$$

$$\Delta_j^n w_{ik}^{nJ+j-1} = g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))w_{i0}^{nJ+j-1} \times g'(w_i^{nJ+j-1} \cdot \xi^j)\xi_k^j + \lambda \frac{f'(w_{ik}^{nJ+j-1})}{2Jf(w_{ik}^{nJ+j-1})^{\frac{1}{2}}} \tag{2.23}$$

where $i = 1, 2, \cdots, q$; $k = 1, 2, \cdots, p$; and η_n is the learning rate in the n-th training epoch.

3. Main Results

For any vector $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$, we write the Euclidean norm of x as $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$. Let $\Omega_0 = \{W : E_W(W) = 0\}$ be the stationary point set of the error function $E(W)$. The following assumptions are imposed for the convergence of the algorithm:

$(A1)$ $|g(t)|$ and $|g'(t)|$ are Lipschitz continuous for $t \in \mathbb{R}$.

$(A2)$ $0 < \eta_n < 1$, $\sum_{n=0}^{\infty} \eta_n < \infty$;

For simplicity, we denote

$$\begin{aligned} C_2 &= \max\{\sup_{t\in\mathbb{R}} |g(t)|, \sup_{t\in\mathbb{R}} |g'(t)|, \sup_{t\in\mathbb{R}} |g''(t)|, \sup_{t\in\mathbb{R}, 1\le j\le J} |g_j'(t)|, \\ &\quad \sup_{t\in\mathbb{R}, 1\le j\le J} |g_j''(t)|\}, \\ C_3 &= \max_{1\le j\le J} \|\xi^j\|, \\ C_4 &= \max\{\sqrt{q}C_2, C_2C_3\}. \end{aligned} \tag{3.1}$$

Theorem 3.1. *Let the error function $E(W)$ be defined by* (2.17)*, W^0 be an arbitrary initial value, and the weight sequence $\{W^n\}$ be generated by the iteration algorithm OGSL1/2* (2.21) *for an arbitrary initial value. Assume the Conditions* $(A1)$ *and* $(A2)$ *are valid, then there exists an unique $W^* \in \Omega_0$ such that*

$$\lim_{n\to\infty} W^n = W^*, \tag{3.2}$$

$$\lim_{n\to\infty} \|E_W(W^n)\| = \|E_W(W^*)\| = 0. \tag{3.3}$$

4. Numerical Examples

To demonstrate the validity of the modified OGSL1/2, we compare it with the OGL2 and OGL1/2 by using two examples: a benchmark problem-parity problem and a nonlinear function regression.

4.1. Example 1: Parity problem

The input set consists of 2^n patterns in n-dimensional space and each pattern is an n-bit binary vector. The target output O^j is equal to 1 if the number of 1 in the pattern is odd, otherwise it is equal to zero. For simplicity of this example, the 3-bit parity problem is considered for the three algorithms with five inputs (including bias), six hidden units (including bias) and one output. All transfer functions are $tansig(\cdot)$. The test has been conducted by selecting the learning rate η and regular factor λ with 0.05 and 0.001, respectively. The training procedure is stopped after 3000 iterations or when the error is less than 1e-6. In order to evaluate the advantage of the OGSL1/2 algorithm, a typical performance of OGL2, OGL1/2 and OGSL1/2 in one of the 10 trials is shown in Fig.2 and 3.

From Fig.2 and 3 it can be seen that the error function $E(W)$ and the norm of the gradient of error function for OGSL1/2 are monotonically decreasing and converge to 0 as predicted by Theorem 3.1, and the results show that the OGSL1/2 are better than OGL2 and OGL1/2. Specifically, OGSL1/2 overcomes the drawback of numerical oscillations at the origin of the OGL1/2.

Table.1 shows the results of 10 trials average errors and norms of gradients of training patterns for each learning algorithm. The Average Number of neurons Eliminated by pruning (ANE in brief) over the 10 tests have been shown in the Table.1 as well. The comparison convincingly supports the finding that OGSL1/2 is more efficient and has more robust sparsity-promoting property than OGL2 and OGL1/2.

Table 1: Simulation results for solving the Parity problem

Learning algorithms	Average error	Norms of gradients	ANE
$OGL2$	0.0071	0.0497	2.6
$OGL1/2$	0.0047	0.0362	8.3
$OGSL1/2$	0.0036	0.0061	13.2

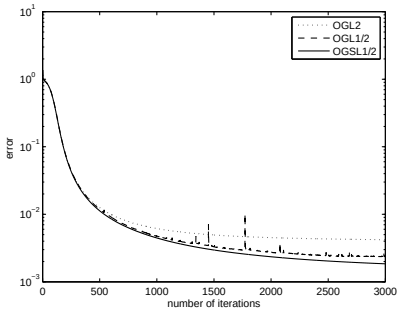

Figure 2: Learning error in Example 1

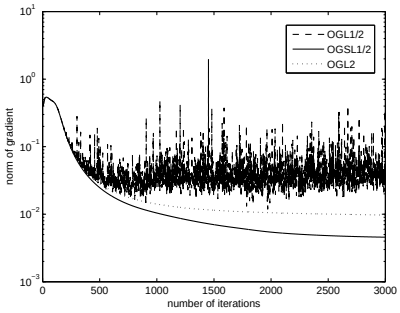

Figure 3: Norm of gradient in Example 1

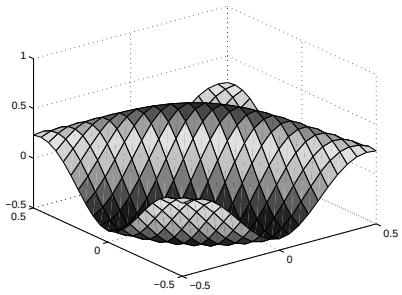

Figure 4: Gabor function in Example 2

4.2. Example 2: Function regression problem

In this example we test the performance of OGL2, OGL1/2 and OGSL1/2 for a multi-dimensional Gabor function:

$$h(x,y) = \frac{1}{2\pi(0.5)^2} exp(-\frac{x^2+y^2}{2\pi(0.5)^2}) cos(2\pi(x+y)) \tag{4.1}$$

In this example, 441 training patterns are selected from an evenly spaced 21×21 grid on the square $0.5 \leq x \leq 0.5$ and $0.5 \leq y \leq 0.5$. 441 testing patterns are selected similarly. The learning rate η is 0.003 and the regularization parameter λ is 0.0003, 10 trials are carried out for each learning algorithm, and all three algorithms are each run 15,000 iterations per trial. The average testing and training error and the ANE across the 10 trails are listed in Table.2. Moreover, in order to compare the approximation quality of the three algorithms, we illustrate a typical performance for one of 10 experiments. Figs.5-7 show the approximation error surfaces for each of the three algorithms, including the OGSL1/2.

From the training and testing results, it can be seen that compared to OGL2 and OGL1/2, OGSL1/2 performs better and yields smaller error and more obvious pruning performance (ANE). Figs.5-7 show that OGSL1/2 has better approximation performance than OGL2 and OGL1/2.

5. Conclusions

A modified smoothing $L_{1/2}$ regularization term learning algorithm has been proposed in this paper. The method has been found to overcome oscillation and its convergence has been proved. In comparison to existing results, this strategy

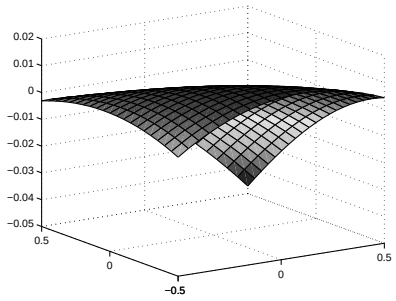

Figure 5: Approximation performance by OGL2

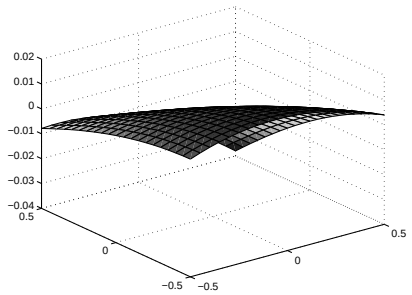

Figure 6: Approximation performance by OGL1/2

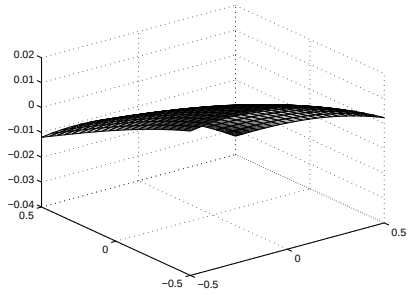

Figure 7: Approximation performance by OGSL1/2

Table 2: Simulation results for identifying Gabor function

Learning algorithms	Average error of training	Average error of test	ANE
$OGL2$	0.0467	0.0642	3.26
$OGL1/2$	0.0292	0.0437	25.69
$OGSL1/2$	0.0203	0.0312	30.14

simplifies the calculation of the gradient of error function and improves the learning efficiency. Therefor, the pruning effect is also more effective for this approach. Moreover, the assumption for the convergence results in our work is a new extension as well. Two supporting numerical examples are provided.

6. Appendix

Before the proofs of Theorem 3.1, the following useful lemmas will be presented.

Lemma 6.1 (Lemma 4.2 in [38])**.** *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*

$$\lim_{m\to\infty} a_m = 0. \tag{6.1}$$

Lemma 6.2 ([39])**.** *Let Y_t, W_t and Z_t be three sequences such that W_t is nonnegative and Y_t is bounded for all t. If*

$$Y_{t+1} \leq Y_t - W_t + Z_t, \ t = 0, 1, \cdots. \tag{6.2}$$

and the series $\sum_{t=0}^{\infty} Z_t$ is convergent, then Y_t converges to a finite value and $\sum_{t=0}^{\infty} W_t < \infty$.

The proof of Theorem 3.1 is divided into two steps.

Step 1. By the assumption (A2), i.e., $\sum_{n=0}^{\infty} \eta_n < \infty$, we can easily get that the sequence $S_n = \eta_0 + \eta_1 + \cdots \eta_{n-1}$ is convergence sequence. By the Cauchy's test for convergence, for $\forall \varepsilon > 0$, there exists a positive integer $N_1 \in \mathbb{N}$, for $\forall n > N_1, \ \forall p \in \mathbb{N}$, we have

$$|S_{n+p} - S_n| = \eta_n + \eta_{n+1} + \cdots + \eta_{n+p-1} < \varepsilon$$

$$|S_{n+p+1} - S_{n+1}| = \eta_{n+1} + \eta_{n+2} + \cdots + \eta_{n+p} < \varepsilon$$

Applying the assumption (A1), there is a constant $A_2 > 0$ such that

$$A_2 = \sup\{|g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))g(w_i^{nJ+j-1} \cdot \xi^j)|\}$$

where ($n \in \mathbb{N},\ j = 1, 2, \cdots, J$)

In addition, for all $x \in \mathbb{R}$, $f(x) \in [\frac{3}{8}a, +\infty)$, $f'(x) \in [-1, 1]$ holds. By the updating formulas (2.21), (2.22) and (2.23), we have

$$\begin{aligned}
&|w_{i0}^{nJ+j} - w_{i0}^{nJ+j-1}| \\
&= \eta_n |\Delta_j^n w_{i0}^{nJ+j-1}| \\
&\le \eta_n (|g_j'(w_0^{nJ+j-1} \cdot G(V^{nJ+j-1}\xi^j))g(w_i^{nJ+j-1} \cdot \xi^j)| \\
&\quad + \lambda |\frac{f'(w_{i0}^{nJ+j-1})}{2Jf(w_{i0}^{nJ+j-1})^{\frac{1}{2}}}|) \\
&\le \eta_n (A_2 + \frac{\lambda}{3a}\sqrt{6a}) \\
&\le \eta_n A_1
\end{aligned} \tag{6.3}$$

where $A_1 = A_2 + \frac{\lambda}{3a}\sqrt{6a}$.

Then

$$\begin{aligned}
&|w_{i0}^{(n+1)J+j} - w_{i0}^{nJ+j}| \\
&\le |w_{i0}^{(n+1)J+j} - w_{i0}^{(n+1)J+j-1}| \\
&\quad + |w_{i0}^{(n+1)J+j-1} - w_{i0}^{(n+1)J+j-2}| \\
&\quad + \cdots + |w_{i0}^{(n+1)J+1} - w_{i0}^{(n+1)J}| \\
&\quad + |w_{i0}^{nJ+J} - w_{i0}^{nJ+J-1}| \\
&\quad + |w_{i0}^{nJ+J-1} - w_{i0}^{nJ+J-2}| \\
&\quad + \cdots + |w_{i0}^{nJ+j+1} - w_{i0}^{nJ+j}| \\
&\le (j\eta_{n+1} + (J-j)\eta_n)A_1
\end{aligned} \tag{6.4}$$

Since

$$\begin{aligned}
&|w_{i0}^{(n+p)J+j} - w_{i0}^{nJ+j}| \\
&\le |w_{i0}^{(n+p)J+j} - w_{i0}^{(n+p-1)J+j}| \\
&\quad + |w_{i0}^{(n+p-1)J+j} - w_{i0}^{(n+p-2)J+j}| \\
&\quad + \cdots + |w_{i0}^{(n+1)J+j} - w_{i0}^{nJ+j}| \\
&\le A_1 j(\eta_{n+p} + \eta_{n+p-1} + \cdots + \eta_{n+1}) \\
&\quad + A_1(J-j)(\eta_{n+p-1} + \eta_{n+p-2} + \cdots + \eta_n) \\
&\le JA_1\varepsilon
\end{aligned} \tag{6.5}$$

Therefore, the weight sequence $\{w_{i0}^{nJ+j}\}$ is a convergence sequence.

By the properties of convergence sequence, $\{w_{i0}^{nJ+j}\}$ must be a bounded sequence, so is $\|w_0^{nJ+j}\|$. Namely, there exists a constant $M_1 > 0$ such that

$$\|w_0^{nJ+j}\| \le M_1$$

where $n = 0, 1, 2, \cdots,\ i = 1, 2, \cdots, q,\ j = 1, 2, \cdots, J$.

Similarly, we can get that $|w_{ik}^{nJ+j}|$ $(n = 0, 1, 2, \cdots,\ i = 1, 2, \cdots, q,\ k = 1, 2, \cdots, p,\ j = 1, 2, \cdots, J)$ is also bounded. So, we obtain the uniform boundness of the weight sequence $\{W^{nJ+j}\}$

$$\|W^{nJ+j}\| \le M,\ n = 0, 1, 2, \cdots,\ j = 1, 2, \cdots, J. \tag{6.6}$$

where $M > 0$ is a suitable constant.

Naturally, there also exists a constant $\bar{M} \ge 0$ such that

$$\|\Delta_k^n w_i^{nJ+k-1}\| \le \bar{M} \tag{6.7}$$

for $k = 1, 2, \cdots J$.

This proof is completed.

Step 2. For convenience, we introduce the following notations:

$$r_i^{n,j} = \Delta_j^n w_i^{nJ+j-1} - \Delta_j^n w_i^{nJ} \tag{6.8}$$

$$G^{n,j} = G(V^n \xi^j), \quad \psi^{n,l,j} = G^{nJ+l-1,j} - G^{nJ,j} \tag{6.9}$$

$$d_i^{n,j} = w_i^{nJ+j-1} - w_i^{nJ}, \quad D_i^{n,j} = W_i^{nJ+l-1} - W_i^{nJ} \tag{6.10}$$

For all $i = 0, 1, 2, \cdots, q$ and $n = 0, 1, 2, \cdots$, we have

$$r_i^{n,1} = 0, \quad d_i^{n,j} = -\eta_n \sum_{k=1}^{j-1} (\Delta_k^n w_i^{nJ} + r_i^{n,k}) \quad j = 1, 2, \cdots, J \tag{6.11}$$

$$\|d_i^{n,j}\| = \eta_n (j-1) \bar{M} \tag{6.12}$$

Indeed

$$\begin{aligned} d_i^{n,j} &= w_i^{nJ+j-1} - w_i^{nJ} \\ &= -\eta_n \Delta_{j-1}^n w_i^{nJ+j-2} - \eta_n \Delta_{j-2}^n w_i^{nJ+j-3} - \cdots \\ &\quad - \eta_n \Delta_2^n w_i^{nJ+1} - \eta_n \Delta_1^n w_i^{nJ} \\ &= -\eta_n \sum_{k=1}^{j-1} \Delta_k^n w_i^{nJ+k-1} \\ &= -\eta_n \sum_{k=1}^{j-1} (\Delta_k^n w_i^{nJ} + r_i^{n,k}) \end{aligned} \tag{6.13}$$

Then, by the error function (2.17), we obtain that

$$\begin{aligned} E(W^{(n+1)J}) = &\sum_{j=1}^{J} g_j(w_0^{(n+1)J} \cdot G(V^{(n+1)J} \xi^j)) \\ &+ \lambda \sum_{i=1}^{q} \sum_{k=0}^{p} f(w_{ik}^{(n+1)J})^{\frac{1}{2}} \end{aligned} \tag{6.14}$$

$$E(W^{nJ}) = \sum_{j=1}^{J} g_j(w_0^{nJ} \cdot G(V^{nJ} \xi^j)) + \lambda \sum_{i=1}^{q} \sum_{k=0}^{p} f(w_{ik}^{nJ})^{\frac{1}{2}} \tag{6.15}$$

Applying the Taylor mean value theorem with Lagrange remainder, we con-

clude that

$$
\begin{aligned}
&E(W^{(n+1)J}) - E(W^{nJ}) \\
&= \sum_{j=1}^{J}(g_j(w_0^{(n+1)J} \cdot G^{(n+1)J,j}) - g_j(w_0^{nJ} \cdot G^{nJ,j})) \\
&+ \lambda \sum_{i=1}^{q}\sum_{k=0}^{p}(f(w_{ik}^{(n+1)J})^{\frac{1}{2}} - f(w_{ik}^{nJ})^{\frac{1}{2}}) \\
&= \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) d_0^{n,J+1} \cdot G^{nJ,j} \\
&+ \lambda \sum_{i=1}^{q}\sum_{k=0}^{p}(f(w_{ik}^{(n+1)J})^{\frac{1}{2}} - f(w_{ik}^{nJ})^{\frac{1}{2}}) \\
&+ \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) w_0^{nJ} \psi^{n,J+1,j} + \delta_1 + \delta_2
\end{aligned}
\tag{6.16}
$$

where $\delta_1 = \frac{1}{2}\sum_{j=1}^{J} g_j''(s_{nJ,j})(w_0^{(n+1)J} \cdot G^{(n+1)J,j} - w_0^{nJ} \cdot G^{nJ,j})^2$, $\delta_2 = \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) d_0^{n,J+1}\psi^{n,J+1,j}$ and $s_{nJ,j} \in \mathbb{R}$ is a constant between $w_0^{nJ} \cdot G^{nJ,j}$ and $w_0^{(n+1)J} \cdot G^{(n+1)J,j}$.

Noting the equation (2.22), we can get that

$$
\begin{aligned}
&\sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) \cdot d_0^{n,J+1} \cdot G^{nJ,j} \\
&= \sum_{i=1}^{q}\sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) g(w_i^{nJ}\xi^j) \cdot d_{i0}^{n,J+1} \\
&= \sum_{i=1}^{q}\sum_{j=1}^{J} \Delta_j^n w_{i0}^{nJ} \cdot d_{i0}^{n,J+1} - \lambda \sum_{i=1}^{q}\sum_{j=1}^{J} \frac{f'(w_{i0}^{nJ}) \cdot d_{i0}^{n,J+1}}{2Jf(w_{i0}^{nJ})^{\frac{1}{2}}}
\end{aligned}
\tag{6.17}
$$

By the Taylor mean value theorem with Lagrange remainder and the equation

(2.23), we have

$$
\begin{aligned}
&\sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) w_0^{nJ} \psi^{n,J+1,j} \\
&= \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) \sum_{i=1}^{q} w_{i0}^{nJ} \cdot [g(w_i^{(n+1)J} \cdot \xi^j) \\
&\quad - g(w_i^{nJ} \cdot \xi^j)] \\
&= \sum_{i=1}^{q} \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) \cdot w_{i0}^{nJ} \cdot g'(w_i^{nJ} \cdot \xi^j) \\
&\quad \times d_i^{n,J+1} \cdot \xi^j + \frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) \cdot w_{i0}^{nJ} \\
&\quad \times g''(t_{i,j,nJ}) \cdot (d_i^{n,J+1} \cdot \xi^j)^2 \\
&= \sum_{k=1}^{p} \sum_{i=1}^{q} \sum_{j=1}^{J} \Delta_j^n w_{ik}^{nJ} \cdot d_{ik}^{n,J+1} \\
&\quad - \lambda \sum_{k=1}^{p} \sum_{i=1}^{q} \sum_{j=1}^{J} \frac{f'(w_{ik}^{nJ}) \cdot d_{ik}^{n,J+1}}{2Jf(w_{ik}^{nJ})^{\frac{1}{2}}} + \delta_3
\end{aligned}
\tag{6.18}
$$

where $t_{i,j,nJ} \in \mathbb{R}$ is between $w_i^{nJ} \cdot \xi^j$ and $w_i^{(n+1)J} \cdot \xi^j$, and $\delta_3 = \frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{J} g_j'(w_0^{nJ} \cdot G^{nJ,j}) w_{0i}^{nJ} g''(t_{i,j,nJ}) (d_i^{n,J+1} \cdot \xi^j)^2$.

Substituting (6.17) and (6.18) into (6.16) and using Taylor mean value theorem

with Lagrange remainder for $F(x) = f(x)^{\frac{1}{2}}$, we have

$$\begin{aligned}
&E(W^{(n+1)J}) - E(W^{nJ}) \\
&= \sum_{k=0}^{p}\sum_{i=1}^{q}\sum_{j=1}^{J}\Delta_j^n w_{ik}^{nJ}\cdot d_{ik}^{n,J+1} + \frac{\lambda}{2}\sum_{k=0}^{p}\sum_{i=1}^{q}F''(t_{i,k,nJ}) \\
&\quad\times (d_{ik}^{n,J+1})^2 + \delta_1 + \delta_2 + \delta_3 \\
&\le -\eta_n\sum_{k=0}^{p}\sum_{i=1}^{q}(\sum_{j=1}^{J}\Delta_j^n w_{ik}^{nJ})^2 - \eta_n\sum_{k=0}^{p}\sum_{i=1}^{q}(\sum_{j=1}^{J}\Delta_j^n w_{ik}^{nJ} \\
&\quad\times\sum_{j=1}^{J}r_{ik}^{n,j}) + M\lambda\sum_{k=0}^{p}\sum_{i=1}^{q}(d_{ik}^{n,J+1})^2 + \delta_1 + \delta_2 + \delta_3 \\
&= -\eta_n\|E_W(W^{nJ})\|^2 - \eta_n\sum_{k=0}^{p}\sum_{i=1}^{q}(\sum_{j=1}^{J}\Delta_j^n w_{ik}^{nJ}\cdot\sum_{j=1}^{J}r_{ik}^{n,j}) \\
&\quad + M\lambda\sum_{k=0}^{p}\sum_{i=1}^{q}(d_{ik}^{n,J+1})^2 + \delta_1 + \delta_2 + \delta_3
\end{aligned} \tag{6.19}$$

where $t_{i,k,nJ} \in \mathbb{R}$ is between w_{ik}^{nJ} and $w_{ik}^{(n+1)J}$, $M = \frac{\sqrt{6}}{4\sqrt{a^3}}$, and $F(x) \equiv (f(x))^{\frac{1}{2}}$. Note that

$$F'(x) = \frac{f'(x)}{2\sqrt{f(x)}}$$

and it is easy to get

$$\begin{aligned}
F''(x) &\le \frac{f''(x)}{2\sqrt{f(x)}} \\
&\le \frac{\sqrt{6}}{2\sqrt{a^3}}
\end{aligned}$$

Since Step 1 shows that $\Delta_j^n w_i^{nJ+j-1}$ ($n \in \mathbb{N}$, $j = 1, 2, \cdots, J$) is bounded, and by the equation (6.7) and (6.13), there exists a suitable constant $C_1 \ge 0$ such that

$$\begin{aligned}
\sum_{k=0}^{p}\sum_{i=1}^{q}(d_{ik}^{n,J+1})^2 &= \sum_{k=0}^{p}\sum_{i=1}^{q}(-\eta_n\sum_{j=1}^{J}\Delta_j^n w_{ik}^{nJ+j-1})^2 \\
&\le (p+1)qJ^2\bar{M}^2\eta_n^2 \\
&= C_1\eta_n^2
\end{aligned} \tag{6.20}$$

where $C_1 = (p+1)qJ^2\bar{M}^2$.

By the Lagrange mean value theorem, $(A1)$, (6.7) and (6.13), we have

$$\begin{aligned}
\|\psi^{n,l,j}\| &= \left\| \begin{pmatrix} g'(\tilde{t}_{1,j,l,n})(w_1^{nJ+l-1} - w_1^{nJ}) \cdot \xi^j \\ \vdots \\ g'(\tilde{t}_{q,j,l,n})(w_q^{nJ+l-1} - w_q^{nJ}) \cdot \xi^j \end{pmatrix} \right\| \\
&\leq (\sup_{t\in\mathbb{R}} |g'(t)| \cdot \max_{1\leq j\leq J} \|\xi^j\|) \sum_{i=1}^{q} \|w_i^{nJ+l-1} - w_i^{nJ}\| \qquad (6.21) \\
&\leq \eta_n (\sup_{t\in\mathbb{R}} |g'(t)| \cdot \max_{1\leq j\leq J} \|\xi^j\|) \sum_{i=1}^{q} \sum_{k=1}^{l-1} \Delta_k^n w_i^{nJ+k-1} \\
&\leq C_4 q(l-1)\bar{M}\eta_n
\end{aligned}$$

and

$$\|G(x)\| \leq \sqrt{q} \sup_{t\in\mathbb{R}} |g(t)| \leq C_4, \quad x \in \mathbb{R}^q \qquad (6.22)$$

where $\tilde{t}_{i,j,l,n} \in \mathbb{R}$ $(1 \leq i \leq q)$ is between $w_i^n \cdot \xi^j$ and $w_i^{n+1} \cdot \xi^j$.

Considering the continuity of $g_j'(t)$, Step 1 and (6.22), we get

$$|g_j'(w_0^{nJ+j-1} \cdot G^{nJ+j-1,j}) - g_j'(w_0^{nJ} \cdot G^{nJ+j-1,j})| \leq LC_4 \|d_0^{n,j}\| \qquad (6.23)$$

$$|g_j'(w_0^{nJ} \cdot G^{nJ+j-1,j}) - g_j'(w_0^{nJ} \cdot G^{nJ,j})| \leq LM_1 \|\psi^{n,j,j}\| \qquad (6.24)$$

By the definition of $r_{i0}^{n,j}$ $(i = 1, 2, \cdots, q)$, we have

$$
\begin{aligned}
|r_{i0}^{n,j}| &= |g_j'(w_{i0}^{nJ+j-1} \cdot G^{nJ+j-1,j})(g(w_i^{nJ+j-1} \cdot \xi^j) \\
&\quad - g(w_i^{nJ} \cdot \xi^j)) + [g_j'(w_0^{nJ+j-1} \cdot G^{nJ+j-1,j}) \\
&\quad - g_j'(w_0^{nJ} \cdot G^{nJ+j-1,j})]g(w_i^{nJ} \cdot \xi^j) \\
&\quad + [g_j'(w_0^{nJ} \cdot G^{nJ+j-1,j}) - g_j'(w_0^{nJ} \cdot G^{nJ,j})] \\
&\quad \times g(w_i^{nJ} \cdot \xi^j) + \frac{\lambda}{2J}[\frac{f'(w_{i0}^{nJ+j-1})}{f(w_{i0}^{nJ+j-1})^{\frac{1}{2}}} - \frac{f'(w_{i0}^{nJ})}{f(w_{i0}^{nJ})^{\frac{1}{2}}}]| \\
&\leq \sup_{t\in\mathbb{R}} |g_j'(t)| \cdot (g(w_i^{nJ+j-1} \cdot \xi^j) - g(w_i^{nJ} \cdot \xi^j)) \\
&\quad + LC_2C_4\|d_0^{n,j}\| + LM_1C_2\|\psi^{n,j,j}\| \\
&\quad + \frac{\lambda}{2J}F''(t_{i0,J,j})|d_{i0}^{n,j}| \\
&\leq C_2(g(w_i^{nJ+j-1} \cdot \xi^j) - g(w_i^{nJ} \cdot \xi^j)) + LC_2C_4\|d_0^{n,j}\| \\
&\quad + LM_1C_2\|\psi^{n,j,j}\| + \frac{\lambda M}{J}|d_{i0}^{n,j}| \\
&\leq C_2g'(\tilde{t}_{i,j,nJ})\|w_i^{nJ+j-1} - w_i^{nJ}\| \cdot \|\xi^j\| + LC_2C_4\|d_0^{n,j}\| \\
&\quad + LM_1C_2\|\psi^{n,j,j}\| + \frac{\lambda M}{J}|d_{i0}^{n,j}| \\
&\leq C_2^2C_3\|d_i^{n,j}\| + LC_2C_4\|d_0^{n,j}\| \\
&\quad + LM_1C_2\|\psi^{n,j,j}\| + \frac{\lambda M}{J}|d_{i0}^{n,j}| \\
&\leq [C_2^2C_3(j-1)\bar{M} + LC_2C_4(j-1)\bar{M} \\
&\quad + LM_1C_2C_4q(j-1)\bar{M} + \frac{\lambda M}{J}(j-1)\bar{M}]\eta_n \\
&= (j-1)\bar{M}[C_2^2C_3 + LC_2C_4(1 + M_1q) + \frac{\lambda M}{J}]\eta_n \\
&= C_5\eta_n
\end{aligned}
\tag{6.25}
$$

where $C_5 = (j-1)\bar{M}[C_2^2C_3 + LC_2C_4(1 + M_1q) + \frac{\lambda M}{J}]$.

Similarly to (6.25), for $i = 1, 2, \cdots, q$, $k = 1, 2, \cdots, p$, there exists a constant $C_6 > 0$ such that

$$
|r_{ik}^{n,j}| \leq C_6\eta_n \tag{6.26}
$$

Indeed

$$
\begin{aligned}
|r_{ik}^{n,j}| &= g_j'(w_0^{nJ+j-1}\cdot G(V^{nJ+j-1}\xi^j))w_{i0}^{nJ+j-1}g'(w_i^{nJ+j-1}\cdot\xi^j)\xi_k^j \\
&\quad - g_j'(w_0^{nJ}\cdot G(V^{nJ}\xi^j))w_{i0}^{nJ}g'(w_i^{nJ}\cdot\xi^j)\xi_k^j \\
&\quad + \lambda\frac{f'(w_{ik}^{nJ+j-1})}{2Jf(w_{ik}^{nJ+j-1})^{\frac{1}{2}}} - \lambda\frac{f'(w_{ik}^{nJ})}{2Jf(w_{ik}^{nJ})^{\frac{1}{2}}} \\
&\le M_1C_3|g_j'(w_0^{nJ+j-1}\cdot G(V^{nJ+j-1}\xi^j))g'(w_i^{nJ+j-1}\cdot\xi^j) \\
&\quad - g_j'(w_0^{nJ}\cdot G(V^{nJ}\xi^j))g'(w_i^{nJ}\cdot\xi^j)| \\
&\quad + |\lambda\frac{f'(w_{ik}^{nJ+j-1})}{2Jf(w_{ik}^{nJ+j-1})^{\frac{1}{2}}} - \lambda\frac{f'(w_{ik}^{nJ})}{2Jf(w_{ik}^{nJ})^{\frac{1}{2}}}| \\
&\le M_1C_3|g_j'(w_{i0}^{nJ+j-1}\cdot G^{nJ+j-1,j})(g'(w_i^{nJ+j-1}\cdot\xi^j) \\
&\quad - g'(w_i^{nJ}\cdot\xi^j)) + [g_j'(w_0^{nJ+j-1}\cdot G^{nJ+j-1,j}) \\
&\quad - g_j'(w_0^{nJ}\cdot G^{nJ+j-1,j})]g'(w_i^{nJ}\cdot\xi^j) \\
&\quad + [g_j'(w_0^{nJ}\cdot G^{nJ+j-1,j}) - g_j'(w_0^{nJ}\cdot G^{nJ,j})] \\
&\quad \times g'(w_i^{nJ}\cdot\xi^j)| + \frac{\lambda}{2J}|\frac{f'(w_{ik}^{nJ+j-1})}{f(w_{ik}^{nJ+j-1})^{\frac{1}{2}}} - \frac{f'(w_{ik}^{nJ})}{f(w_{ik}^{nJ})^{\frac{1}{2}}}| \\
&\le M_1C_3(\sup_{t\in\mathbb{R}}|g_j'(t)|\cdot(g'(w_i^{nJ+j-1}\cdot\xi^j) - g'(w_i^{nJ}\cdot\xi^j)) \\
&\quad + LC_2C_4\|d_0^{n,j}\| + LM_1C_2\|\psi^{n,j,j}\|) \\
&\quad + \frac{\lambda}{2J}F''(t_{ik,J,j})|d_{ik}^{n,j}| \\
&\le M_1C_3(C_2(g'(w_i^{nJ+j-1}\cdot\xi^j) - g'(w_i^{nJ}\cdot\xi^j)) \\
&\quad + LC_2C_4\|d_0^{n,j}\| + LM_1C_2\|\psi^{n,j,j}\|) + \frac{\lambda M}{J}|d_{ik}^{n,j}| \\
&\le M_1C_3(C_2\lambda\|w_i^{nJ+j-1} - w_i^{nJ}\|\cdot\|\xi^j\| + LC_2C_4\|d_0^{n,j}\| \\
&\quad + LM_1C_2\|\psi^{n,j,j}\|) + \frac{\lambda M}{J}|d_{ik}^{n,j}| \\
&\le M_1C_3(\lambda C_2C_3\|d_i^{n,j}\| + LC_2C_4\|d_0^{n,j}\| \\
&\quad + LM_1C_2\|\psi^{n,j,j}\|) + \frac{\lambda M}{J}|d_{ik}^{n,j}| \\
&\le M_1C_3[\lambda C_2C_3(j-1)\bar{M} + LC_2C_4(j-1)\bar{M} \\
&\quad + LM_1C_2C_4q(j-1)\bar{M} + \frac{\lambda M}{J}(j-1)\bar{M}]\eta_n \\
&= (j-1)\bar{M}M_1C_3[\lambda C_2C_3 + LC_2C_4(1+M_1q) + \frac{\lambda M}{J}]\eta_n \\
&= C_6\eta_n
\end{aligned}
\tag{6.27}
$$

where $C_6 = (j-1)\bar{M}M_1C_3[\lambda C_2C_3 + LC_2C_4(1+M_1q) + \frac{\lambda M}{J}]$.

From the above two inequalities and (6.7), we have

$$\begin{aligned} & -\eta_n \sum_{k=0}^{p}\sum_{i=1}^{q}(\sum_{j=1}^{J}\Delta_j^n w_{ik}^{nJ}\cdot\sum_{j=1}^{J} r_{ik}^{n,j}) \\ & \leq \sum_{k=0}^{p}\sum_{i=1}^{q}(J\bar{M}\cdot\max\{C_5,C_6\})\eta_n^2 \\ & = (p+1)qJ\bar{M}\cdot\max\{C_5,C_6\}\eta_n^2 \\ & = C_7\eta_n^2 \end{aligned} \tag{6.28}$$

where $C_7 = (p+1)qJ\bar{M}\cdot\max\{C_5,C_6\}$

It follows from (6.20), (6.21) and the Cauchy-Schwartz inequality that

$$\begin{aligned} |\delta_2| &= |\sum_{j=1}^{J} g_j'(w_0^{nJ}\cdot G^{nJ,j})d_0^{n,J+1}\psi^{n,J+1,j}| \\ &\leq \frac{C_2}{2}\sum_{j=1}^{J}(\|d_0^{n,J+1}\|^2 + \|\psi^{n,J+1,j}\|^2) \\ &\leq \frac{1}{2}JC_2(1+C_4^2)\sum_{k=0}^{p}\sum_{i=1}^{q}(d_{ik}^{n,J+1})^2 \\ &\leq C_8\eta_n^2 \end{aligned} \tag{6.29}$$

where $C_8 = \frac{1}{2}JC_1C_2(1+C_4^2)$.

Similarly, we have

$$
\begin{aligned}
|\delta_1| &= |\frac{1}{2}\sum_{j=1}^{J} g_j''(s_{nJ,j})(w_0^{(n+1)J}\cdot G^{(n+1)J,j} - w_0^{nJ}\cdot G^{nJ,j})^2| \\
&\leq \frac{C_2}{2}\sum_{j=1}^{J}(d_0^{n,J+1}\cdot G^{(n+1)J,j} + w_0^{nJ}\cdot \psi^{n,J+1,j})^2 \\
&\leq \frac{C_2}{2}\sum_{j=1}^{J}(C_4\|d_0^{n,J+1}\| + M_1\|\psi^{n,J+1,j}\|)^2 \\
&\leq C_9\sum_{j=1}^{J}(\|d_0^{n,J+1}\|^2 + C_4^2\sum_{i=1}^{q}\|d_i^{n,J+1}\|^2)) \\
&\leq JC_9(1+C_4^2)\sum_{k=0}^{p}\sum_{i=1}^{q}(d_{ik}^{n,J+1})^2 \\
&\leq C_{10}\eta_n^2
\end{aligned}
\tag{6.30}
$$

where $C_9 = C_2\max\{C_4^2, M_1^2\}$, $C_{10} = JC_1C_9(1+C_4^2)$.

Similarly, we also conclude that

$$
\begin{aligned}
|\delta_3| &= |\frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{J} g_j'(w_0^{nJ}\cdot G^{nJ,j})w_{0i}^{nJ}g''(t_{i,j,nJ})(d_i^{n,J+1}\cdot\xi^j)^2| \\
&\leq \frac{1}{2}JC_2^2C_3^2M_1\sum_{k=0}^{p}\sum_{i=1}^{q}(d_{ik}^{n,J+1})^2 \\
&\leq C_{11}\eta_n^2
\end{aligned}
\tag{6.31}
$$

where $C_{11} = \frac{1}{2}JC_2^2C_3^2M_1$

From (6.19)-(6.31), we can obtain

$$
E(W^{(n+1)J}) \leq E(W^{nJ}) - \eta_n\|E_W(W^{nJ})\|^2 + M_0\eta_n^2 \tag{6.32}
$$

where $M_0 = C_7 + M\lambda C_1 + C_{10} + C_8 + C_{11} \geq 0$.

Indeed

$$\begin{aligned}
&E(W^{(n+1)J}) - E(W^{nJ}) \\
&= -\eta_n \|E_W(W^{nJ})\|^2 - \eta_n \sum_{k=0}^{p} \sum_{i=1}^{q} (\sum_{j=1}^{J} \Delta_j^n w_{ik}^{nJ} \cdot \sum_{j=1}^{J} r_{ik}^{n,j}) \\
&\quad + M\lambda \sum_{k=0}^{p} \sum_{i=1}^{q} (d_{ik}^{n,J+1})^2 + \delta_1 + \delta_2 + \delta_3 \\
&\leq -\eta_n \|E_W(W^{nJ})\|^2 + C_7 \eta_n^2 + M\lambda C_1 \eta_n^2 \\
&\quad + C_8 \eta_n^2 + C_{10} \eta_n^2 + C_{11} \eta_n^2 \\
&\leq -\eta_n \|E_W(W^{nJ})\|^2 + (C_7 + M\lambda C_1 + C_8 + C_{10} + C_{11}) \eta_n^2 \\
&= -\eta_n \|E_W(W^{nJ})\|^2 + M_0 \eta_n^2
\end{aligned} \tag{6.33}$$

By the assumption, Lemma 6.2 and the above equation, we can see that

$$\sum_{n=0}^{\infty} \eta_n \|E_W(W^{nJ})\|^2 = \eta_n \sum_{k=0}^{p} \sum_{i=1}^{q} [E_{w_{ik}}(W^{nJ})]^2 < \infty \tag{6.34}$$

By the comparison discriminant method of the series, we get

$$\sum_{n=0}^{\infty} \eta_n [E_{w_{ik}}(W^{nJ})]^2 < \infty \tag{6.35}$$

It follows from (2.19), (2.20), (2.22) and (2.23), and the similar estimation with (6.27), there exists a suitable constant C_{12} such that

$$\begin{aligned}
&|E_{w_{ik}}(W^{(n+1)J}) - E_{w_{ik}}(W^{nJ})| \\
&= \sum_{j=1}^{J} |r_{ik}^{n,J+1}| \\
&\leq J \max\{C_5, C_6\} \eta_n \\
&= C_{12} \eta_n
\end{aligned} \tag{6.36}$$

where $C_{12} = J \max\{C_5, C_6\}$.

Combining the above two inequations (6.35), (6.36) and using Lemma 6.1 immediately gives

$$\lim_{n \to \infty} E_{w_{ik}}(W^{nJ}) = 0 \tag{6.37}$$

Considering the assumption (A2), it is easy to get

$$\lim_{n\to\infty} \eta_n = 0 \tag{6.38}$$

By the equation (6.36), we obtain

$$\begin{aligned} |E_{w_{ik}}(W^{nJ+j})| &\le \sum_{j=1}^{J} |r_{ik}^{n,j+1}| + |E_{w_{ik}}(W^{nJ})| \\ &\le C_{12}\eta_n + |E_{w_{ik}}(W^{nJ})| \end{aligned} \tag{6.39}$$

thus, for $n = 0, 1, 2, \cdots,\ i = 1, 2, \cdots, q,\ k = 0, 1, 2, \cdots, p,\ j = 1, 2, \cdots, J$

$$\lim_{n\to\infty} E_{w_{ik}}(W^{nJ+j}) = 0 \tag{6.40}$$

Hence

$$\lim_{n\to\infty} \|E_W(W^{nJ+j})\| = 0. \tag{6.41}$$

According to the assumption (A1), it indicates that $E_W(W)$ is continuous. From the proof of Step 1, we know that the sequence $\{W^{nJ+j}\}$ is a convergence sequence. Let $\lim_{n\to\infty} W^{nJ+j} = W^*$, then $\lim_{n\to\infty} E_W(W^{nJ+j}) = E_W(W^*) = 0$.

Hence,

$$\lim_{n\to\infty} W^{nJ+j} = W^*, W^* \in \Omega_0. \tag{6.42}$$

This proof is completed.

References

[1] P.J. Werbos, Bryond Regression: New Tools for Prediction and Analysis in the Behavioral Science. Ph.D. Thsis, Harvard University, 1974.

[2] S. Haykin, Neural Networks: A Comprehensive Foundation,2nd edition, Tsinghua University Press, PrenticeHall, Beijing, 2001.

[3] J.M. Zurada, Introduction to Artificial Neural Systems. West Publishing Company, 1992.

[4] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533-536. Reprinted in Anderson and Rosenberg (1988).

[5] D. Saad, On-line learning in neural networks, Cambridge, England, New York: Cambridge University Press, 1998.

[6] T. Heskes, W. Wiegerinck, A Theoretical Comparison of Batch-Mode, On-Line, Cyclic, and Almost-Cyclic Learning, IEEE Transactions on Neural Networks 7 (1996) 919-925.

[7] J.A. Farrell, On performance evaluation in on-line approxmation for control, IEEE Transactions on Neural Networks 9 (1998) 1001-1007.

[8] F. Beaufays, E.A. Wan, Relating real-time backpropagation and backpropagation through time: An application of flow graph inter-reciprocity, Neural Computation 6 (1994) 296-306.

[9] S.W. Ellacott, The numerical analysis approach of neural networks, North-Holland Mathematical Library 51 (1993) 103-137.

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

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

[12] J. Wang, W. Wu, J.M. Zurada, Deterministic convergence of conjugate gradient method for feedforward neural networks, Neurocomputing 74 (2011) 2368-2376.

[13] F. Han, D.S. Huang, Improved extreme learning machine for function approximation by encoding a priori information, Neurocomputing 69 (2006) 2369-2373.

[14] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (2006) 489-501.

[15] W. Wu, Y.S. Xu, Deterministiic convergeence of online gradient method for neural networks, Journal of Computational and Applied Mathematics 144 (2002) 335-347.

[16] W. Wu, G.R. Feng, X.Li, Trainging multilayer perceptrons via miniization of sun of ridge functions, Advances in Computational Mathematics 17 (2002) 331-347.

[17] H.S. Zhang, W. Wu, Boundedness and convergence of online gradient method with penalty for linear output feedforward neural networks, Neural Process Lett 29 (2009) 205-212.

[18] G.E. Hinton, Connectionist learning procedures, Artificial Intelligence 40 (1989) 185-234.

[19] S. Geman, E. Bienenstock, R.Doursat, Neural networks and the bias/variance dilemma, Neural Computation 4 (1992) 1-58.

[20] M. Ishikawa, Structural learning with forgetting, Neural Networks 9 (1996) 509-521.

[21] P.L. Bartlett, For valid generalization the size of the weights is more important than the size of the network, In Advances in Neural Information Processing Systems 9 (1997) 134-140.

[22] R. Reed, Pruning algorithms - a survey, IEEE Transactions on Neural Networks 4 (1993) 740-747.

[23] R. Setiono, A penalty-function approach for pruning feedforward neural networks, Neural Computation 9 (1997) 185-204.

[24] H.M. Shao, W. Wu, L.J. Liu, Convergnece of online gradient method with a penalty for BP neural networks, Communications in Mathematical Research 26 (2010) 67-75.

[25] A.S. Weigend, D.E. Rumelhart, B.A. Huberman, Generalization by weightelimination applied to currency exchange rate prediction, IJCNN-91-Seattle International Joint Conference on Neural Networks 1 (1991) 837-841.

[26] J.E. Moody, T.S. Rognvaldsson, Smoothing regularizers for projective basis function networks, Advances in Neural Information Processing Systems, (1996) 585-591.

[27] J. Kong, W. Wu, Online gradient methods with a punishing term for neural networks, Northeastern Mathematical Journal 3 (2001) 371-378.

[28] K. Saito, R. Nakano, Second-order learning algorithm with squared penalty term, Neural Computation 12 (2000) 709-729.

[29] H. Akaike, Information theory and an extension of the maximum likelihood principle, In: The Second International Symposium on Information Theory, Akademiai Kiado, Budapest (1973) 267-281.

[30] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (1978) 461-464.

[31] C.L. Mallows, Some comments on Cp, Technometrics 15 (1973) 661-675.

[32] Z.B Xu, H. Zhang, Y. Wang, X.Y. Chang, Y, Liang, L1/2 regularizer, Science China Informaction Sciences 53 (2010) 1159-1169.

[33] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of Royal Statistical Society. Series B(Methodological) 58 (1996) 267-288.

[34] D.L. Donoho, X.M. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Transactions on Information Theory 47 (2001) 2845-2862.

[35] D.L. Donoho, M. Elad, Maximal sparsity representation via L1 minimization, Proceedings of National Academy of Sciences 100 (2003) 2197-2202.

[36] S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1998) 33-61.

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

[38] W. Wu, J. Wang, M.S. Cheng, Z.X. Li, Convergence analysis of online gradient method for BP neural networks, Neural Networks 24 (2011) 91-98.

[39] Z.B. Xu, R. Zhang, W.F. Jing, When Does Online BP Training Converge? IEEE Transactions on Neural Networks 20 (2009) 1529-1539.