Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2012, Article ID 456919, 25 pages
doi:10.1155/2012/456919

*Research Article*

# A Smoothing Interval Neural Network

**Dakun Yang and Wei Wu**

*School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*

Correspondence should be addressed to Wei Wu, wuweiw@dlut.edu.cn

Received 23 May 2012; Accepted 18 October 2012

Academic Editor: Daniele Fournier-Prunaret

In many applications, it is natural to use interval data to describe various kinds of uncertainties. This paper is concerned with an interval neural network with a hidden layer. For the original interval neural network, it might cause oscillation in the learning procedure as indicated in our numerical experiments. In this paper, a smoothing interval neural network is proposed to prevent the weights oscillation during the learning procedure. Here, by smoothing we mean that, in a neighborhood of the origin, we replace the absolute values of the weights by a smooth function of the weights in the hidden layer and output layer. The convergence of a gradient algorithm for training the smoothing interval neural network is proved. Supporting numerical experiments are provided.

## 1. Introduction

In the last two decades artificial neural networks have been successfully applied to various domains, including pattern recognition [1], forecasting [2, 3], and data mining [4, 5]. One of the most widely used neural networks is the feedforward neural network with the well-known error backpropagation learning algorithm. But in most neural network architectures, input variables and the predicted results are represented in the form of single point value, not in the form of intervals. However, in real-life situations, available information is often uncertain, imprecise, and incomplete, which can be represented by fuzzy data, a generalization of interval data. So in many applications it is more natural to treat the input variables and the predicted results in the form of intervals than a set of single-point value.

Since multilayer feedforward neural networks have high capability as a universal approximator of nonlinear mappings [6–8], some methods via neural networks for handling interval data have been proposed. For instance, in [9], the BP algorithm [10, 11] was extended to the case of interval input vectors. In [12], the author proposed a new extension of backpropagation by using interval arithmetic which called Interval Arithmetic Backpropagation (IABP). This new algorithm permits the use of training samples and targets which can be indistinctly points and intervals. In [13], the author proposed a new model of multilayer perceptron based on interval arithmetic that facilitates handling input and output interval data, where weights and biases are single valued and not interval valued.

However, weights oscillation phenomena during the learning procedure were observed in our numerical experiments for these interval neural networks models. In order to prevent the weights oscillation, a smoothing interval neuron is proposed in this paper. Here, by smoothing we mean that, in the activation function and in a neighborhood of the origin, we replace the absolute values of the weights by a smooth function of the weights. Gradient algorithms [14–17] are applied to train the smoothing interval neural network. The weak and strong convergence theorems of the algorithms are proved. Supporting numerical results are provided.

The remainder of this paper is organized as follows. Some basic notations of interval analysis are described in Section 2. The traditional interval neural network is introduced in Section 3. Section 4 is devoted to our smoothing interval neural network and the gradient algorithm. The convergence results of the gradient learning algorithm are shown in Section 5. Supporting numerical experiments are provided in Section 6. The appendix is devoted to the proof of the theorem.

## 2. Interval Arithmetic

Interval arithmetic as a tool appeared in numerical computing in late 1950s. Then the interval mathematic is a theory introduced by Moore [18] and Sunaga [19] in order to give control of errors in numeric computations. Fundamentals used in this paper are described below.

Let us denote the intervals by uppercase letters such as $A$ and the real numbers by lowercase letters such as $a$. An interval can be represented by its lower bounds $L$ and upper bounds $U$ as $A = [a^L, a^U]$, or equivalently by its midpoint $C$ and radius $R$ as $A = \langle a^C, a^R \rangle$, where

$$\begin{aligned} a^C &= \frac{a^L + a^U}{2}, \\ a^R &= \frac{a^U - a^L}{2}. \end{aligned} \tag{2.1}$$

For intervals $A = [a^L, a^U]$ and $B = [b^L, b^U]$, the basic interval operations are defined by

$$\begin{aligned} A + B &= \left[a^L + b^L, a^U + b^U\right], \\ A - B &= \left[a^L - b^U, a^U - b^L\right], \\ k \cdot A &= \begin{cases} \left[k \cdot a^L, k \cdot a^U\right], & k > 0, \\ \left[k \cdot a^U, k \cdot a^L\right], & k < 0, \end{cases} \end{aligned} \tag{2.2}$$

where $k$ is a constant.

If $f$ is an increasing function, then the interval output is given by

$$f(A) = \left[f\left(a^L\right), f\left(a^U\right)\right]. \tag{2.3}$$

In this paper, we use the following weighted Euclidean distance for a pair of intervals $A$ and $B$

$$d(A,B) = \beta\left(a^C - b^C\right)^2 + (1-\beta)\left(a^R - b^R\right)^2, \quad \beta \in (0,1). \tag{2.4}$$

The parameter $\beta \in [0,1]$ facilitates giving more importance to the prediction of the output centres or to the prediction of the radii. For $\beta = 1$ learning concentrates on the prediction of the output interval centre and no importance is given to the prediction of its radius. For $\beta = 0.5$ both predictions (centres and radii) have the same weights in the objective function. For our purpose, we assume $\beta \in (0,1)$.

## 3. Interval Neural Network

In this paper, we consider an interval neural network with three layers, where the input and output are interval value, the weights are real value. The numbers of neurons for the input, hidden and output layers are $N, M, 1$, respectively. Let $W_m = (w_{m1}, w_{m2}, \ldots, w_{mN})^T \in \mathbb{R}^N$, $m = 1, 2, \ldots, M$ be the weight matrix connecting the input and the hidden layers. The weight vector connecting the hidden and the output layers is denoted by $W_0 = (w_{0,1}, w_{0,2}, \ldots, w_{0,M})^T \in \mathbb{R}^M$. To simplify the presentation, we write $W = (W_0^T, W_1^T, \ldots, W_M^T)^T \in \mathbb{R}^{NM+M}$. In the interval neural network, a nonlinear activation function $f(x)$ is used in the hidden layer, and a linear activation function in the output layer.

For an arbitrary interval-valued input $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, where $X_i = \langle x_i^C, x_i^R \rangle$, $i = 1, 2, \ldots, N$, as the weights of the proposed structure are real value, this linear combination results in a interval given by

$$S_m = \sum_{i=1}^{N} w_{mi} X_i = \left\langle s_m^C, s_m^R \right\rangle = \left\langle \sum_{i=1}^{N} w_{mi} x_i^C, \sum_{i=1}^{N} |w_{mi}| x_i^R \right\rangle. \tag{3.1}$$

Then the output of the interval neuron in the hidden layer is given by

$$\begin{aligned} H_m = f(S_m) &= \left[f\left(s_m^C - s_m^R\right), f\left(s_m^C + s_m^R\right)\right] = \left\langle h_m^C, h_m^R \right\rangle \\ &= \left\langle \frac{f\left(s^C - s^R\right) + f\left(s^C + s^R\right)}{2}, \frac{f\left(s^C + s^R\right) - f\left(s^C - s^R\right)}{2} \right\rangle. \end{aligned} \tag{3.2}$$

Finally, the output of the interval neuron in the output layer is given by

$$Y = \left\langle y^C, y^R \right\rangle, \tag{3.3}$$

$$y^C = \sum_{m=1}^{M} w_{0m} h_m^C, \tag{3.4}$$

$$y^R = \sum_{m=1}^{M} |w_{0m}| h_m^R. \tag{3.5}$$

# 4. Smoothing Interval Neural Network

## *4.1. Smoothing Interval Neural Network Structure*

As revealed in the numerical experiment below in this paper, there appear weights oscillation phenomena during the learning procedure for the original interval neural network presented in the last section. In order to prevent the weights oscillation, we propose a smoothing interval neural network by replacing $|w_{mi}|$ and $|w_{0m}|$ with a smooth function $\varphi(w_{mi})$ and $\varphi(w_{0m})$ in (3.1) and (3.5). Then, the output of the smoothing interval neuron in the hidden layer is defined as

$$S_m = \sum_{i=1}^{N} w_{mi} X_i = \left\langle \sum_{i=1}^{N} w_{mi} x_i^C, \sum_{i=1}^{N} \varphi(w_{mi}) x_i^R \right\rangle, \tag{4.1}$$

$$\begin{aligned} H_m = f(S_m) &= \left[ f\left(s_m^C - s_m^R\right), f\left(s_m^C + s_m^R\right) \right] = \left\langle h_m^C, h_m^R \right\rangle \\ &= \left\langle \frac{f\left(s_m^C - s_m^R\right) + f\left(s_m^C + s_m^R\right)}{2}, \frac{f\left(s_m^C + s_m^R\right) - f\left(s_m^C - s_m^R\right)}{2} \right\rangle. \end{aligned} \tag{4.2}$$

The output of the smoothing interval neuron in the output layer is given by

$$\begin{aligned} y^C &= \sum_{m=1}^{M} w_{0m} h_m^C, \\ y^R &= \sum_{m=1}^{M} \varphi(w_{0m}) h_m^R. \end{aligned} \tag{4.3}$$

For our purpose, $\varphi(x)$ can be chosen as any smooth function that approximates $|x|$ near the origin. For definiteness and simplicity, we choose $\varphi(x)$ as a polynomial function:

$$\varphi(x) = \begin{cases} -x, & x \le -\mu, \\ \widehat{\varphi}(x), & -\mu < x < \mu, \\ x, & x \ge \mu, \end{cases} \tag{4.4}$$

where $\mu > 0$ is a small constant and

$$\widehat{\varphi}(x) = -\frac{1}{8\mu^3}x^4 + \frac{3}{4\mu}x^2 + \frac{3}{8}\mu. \tag{4.5}$$

We observe that the above defined $\varphi(x)$ is a convex function in $C^2(R)$, and it is identical to the absolute value function $|x|$ outside the zero neighborhood $(-\mu, \mu)$.

### *4.2. Gradient Algorithm of the Smoothing Interval Neural Network*

Suppose that we are supplied with a training sample set $\{X_j, O_j\}_{j=1}^{J}$, where $X_j$'s and $O_j$'s are input and ideal output samples, respectively, as follows: $X_j = (X_{1j}, X_{2j}, \ldots, X_{Nj})^T$, $X_{ij} = [x_{ij}^L, x_{ij}^U] = \langle x_{ij}^C, x_{ij}^R \rangle$, $i = 1, 2, \ldots, N$, $O_j = [o_j^L, o_j^U] = \langle o_j^C, o_j^R \rangle$. Our task is to find the weights $W = (W_0^T, W_1^T, \ldots, W_M^T)^T$ such that

$$O_j = Y(X_j), \quad j = 1, 2, \ldots, J. \tag{4.6}$$

But usually, the weight $W = (W_0^T, W_1^T, \ldots, W_M^T)^T$ satisfying (4.6) does not exit and, instead, the aim of the network learning is to choose the weight $W$ to minimize an error function of the smoothing interval neural network. By (2.4), a simple and typical error function is the quadratic error function:

$$E(W) = \frac{1}{2}\sum_{j=1}^{J}\left(\beta\left(o_j^C - y_j^C\right)^2 + (1-\beta)\left(o_j^R - y_j^R\right)^2\right). \tag{4.7}$$

Let us denote $f_j^C(t) = (1/2)(o_j^C - t_j^C)^2$, $f_j^R(t) = (1/2)(o_j^R - t_j^R)^2$, $j = 1, 2, \ldots, J$, $t \in \mathbb{R}$, then the error function (4.7) is rewritten as

$$E(W) = \sum_{j=1}^{J}\left(\beta f_j^C(y) + (1-\beta) f_j^R(y)\right). \tag{4.8}$$

Now, we introduce the gradient algorithm [15, 16] for the smoothing interval neural network. The gradient of the error function $E(W)$ with respect to $W_0$ is given by

$$\begin{aligned} \frac{\partial E(W)}{\partial W_0} &= \sum_{j=1}^{J}\left(\beta\left(y_j^C - o_j^C\right)\frac{\partial y_j^C}{\partial W_0} + (1-\beta)\left(y_j^R - o_j^R\right)\frac{\partial y_j^R}{\partial W_0}\right) \\ &= \sum_{j=1}^{J}\left(\beta f_j^{\prime C}(y)\frac{\partial y_j^C}{\partial W_0} + (1-\beta) f_j^{\prime R}(y)\frac{\partial y_j^R}{\partial W_0}\right), \end{aligned} \tag{4.9}$$

where

$$\begin{aligned} \frac{\partial y_j^C}{\partial W_0} &= h_j^C, \\ \frac{\partial y_j^R}{\partial W_0} &= \varphi'(W_0) h_j^R. \end{aligned} \tag{4.10}$$

The gradient of the error function $E(W)$ with respect to $W_m,\ m = 1, 2, \ldots, M$ is given by

$$\begin{aligned} \frac{\partial E(W)}{\partial W_m} &= \sum_{j=1}^{J} \left( \beta \left( y_j^C - o_j^C \right) \frac{\partial y_j^C}{\partial h_{jm}^C} \frac{\partial h_{jm}^C}{\partial W_m} + (1-\beta) \left( y_j^R - o_j^R \right) \frac{\partial y_j^R}{\partial h_{jm}^R} \frac{\partial h_{jm}^R}{\partial W_m} \right) \\ &= \sum_{j=1}^{J} \left( \beta f_j^{'C}(y) \frac{\partial y_j^C}{\partial h_{jm}^C} \frac{\partial h_{jm}^C}{\partial W_m} + (1-\beta) f_j^{'R}(y) \frac{\partial y_j^R}{\partial h_{jm}^R} \frac{\partial h_{jm}^R}{\partial W_m} \right), \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} \frac{\partial y_j^C}{\partial h_{jm}^C} &= w_{0m}, \\ \frac{\partial y_j^R}{\partial h_{jm}^R} &= \varphi(w_{0m}), \\ \frac{\partial h_{jm}^C}{\partial W_m} &= \frac{f'\left(s_{jm}^C - s_{jm}^R\right)\left(x_j^C - \varphi'(W_m) x_j^R\right)}{2} + \frac{f'\left(s_{jm}^C + s_{jm}^R\right)\left(x_j^C + \varphi'(W_m) x_j^R\right)}{2}, \\ \frac{\partial h_{jm}^R}{\partial W_m} &= \frac{f'\left(s_{jm}^C + s_{jm}^R\right)\left(x_j^C + \varphi'(W_m) x_j^R\right)}{2} - \frac{f'\left(s_{jm}^C - s_{jm}^R\right)\left(x_j^C - \varphi'(W_m) x_j^R\right)}{2}. \end{aligned} \tag{4.12}$$

In the learning procedure, the weights $W$ are iteratively refined as follows:

$$W^{k+1} = W^k + \Delta W^k, \tag{4.13}$$

where

$$\Delta W^k = -\eta \frac{\partial E(W^k)}{\partial W}, \tag{4.14}$$

where $\eta > 0$ a constant learning rate and $k = 1, 2, \ldots$.

## 5. Convergence Theorem for SINN

For any $\mathbf{x} \in \mathbb{R}^n$, its Euclidean norm is $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$. Let $\Omega_0 = \{W \in \Omega : E_W(W) = 0\}$ be the stationary point set of the error function $E(W)$, where $\Omega \subset \mathbb{R}^{NM+M}$ is a bounded region satisfying $(A2)$ below. Let $\Omega_{0,s} \subset \mathbb{R}$ be the projection of $\Omega_0$ onto the $s$th coordinate axis, that is,

$$\Omega_{0,s} = \left\{ w_s \in \mathbb{R} : W = (w_1, \ldots, w_s, \ldots, w_{NM+M})^T \in \Omega_0 \right\}, \tag{5.1}$$

for $s = 1, 2, \ldots, NM + M$. To analyze the convergence of the algorithm, we need the following assumptions.

$(A1)$ $|f(t)|$, $|f'(t)|$, $|f''(t)|$ are uniformly bounded for $t \in R$.

$(A2)$ There exists a bounded region $\Omega \subset \mathbb{R}^n$ such that $\{W^k\} \subset \Omega$ $(k \in \mathbb{N})$.

$(A3)$ The learning rate $\eta$ is small enough such that (A.10) below is valid.

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

Now we are ready to present one convergence theorem of the learning algorithms. Its proof is given in the appendix later on.

**Theorem 5.1.** *Let the error function $E(W)$ be defined by* (4.7)*, and the weight sequence $\{W^k\}$ be generated by the learning procedure* (4.13) *and* (4.14) *for smoothing interval neuron with $W^0$ being an arbitrary initial guess. If Assumptions* $(A1)$, $(A2)$, *and* $(A3)$ *are valid, then we have*

$$E\left(W^{k+1}\right) \leq E\left(W^k\right), \tag{5.2}$$

$$\lim_{k \to \infty} \left\| E_W\left(W^k\right) \right\| = 0. \tag{5.3}$$

*Furthermore, if Assumption* $(A4)$ *also holds, there exists a point $W^* \in \Omega_0$ such that*

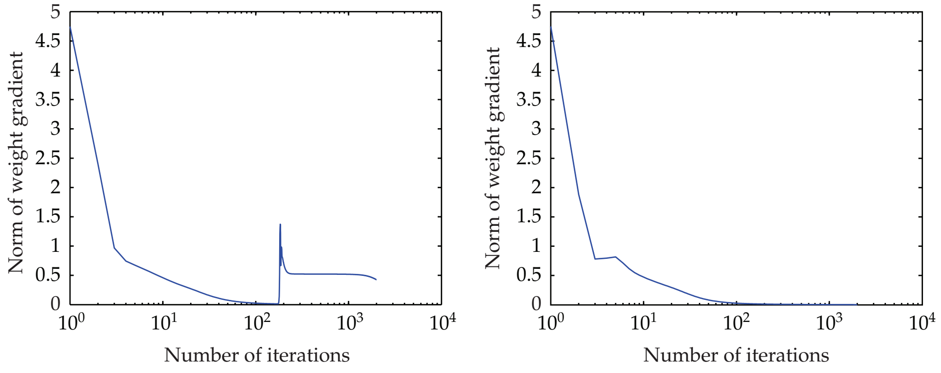
$$\lim_{k \to \infty} W^k = W^*. \tag{5.4}$$

## 6. Numerical Experiment

We compare the performances of the interval neural network and the smoothing interval neural network by approximating a simple interval function

$$Y = 0.01 \times (X + 11)^2. \tag{6.1}$$

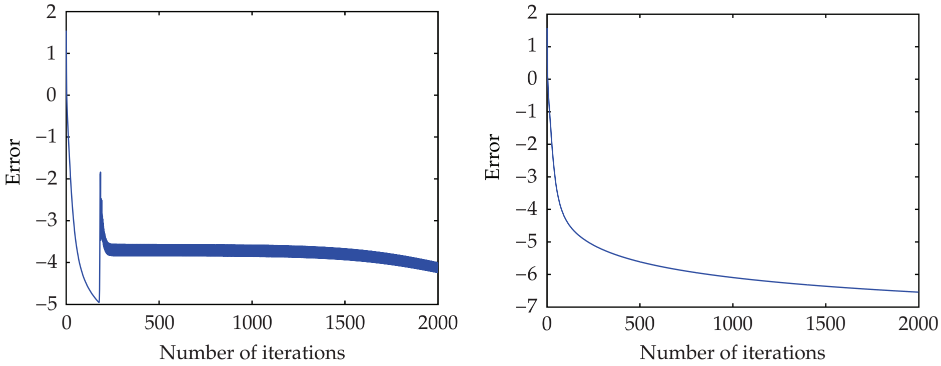
In this example, the training set contains five training samples. Their midpoints are all 0 and their radii are (0.8552 2.6248 8.0101 0.2922 9.2885), respectively. The corresponding outputs of the samples are $Y = \langle y^C, y^R \rangle = 0.01 \times (X + 11)^2$.



(a) Interval neural network

(b) Smoothing interval neural network

**Figure 1:** Norm of gradient of the interval neural network and the smoothing interval neuron in the training.



(a) Value of $D$ for interval neural network

(b) Value of $D$ for smoothing interval neural network

**Figure 2:** Values of the error function $D$ for the interval neural network and the smoothing neural network.

For the above two interval neural networks, the error function $E(W)$ is defined as in (4.7). But in order to see the error more clearly in the figures, we will also use the error $D$ defined by

$$D = \ln E = \ln\left(\frac{1}{2}\sum_{j=1}^{J}\left(\beta\left(o_j^C - y_j^C\right)^2 + (1-\beta)\left(o_j^R - y_j^R\right)^2\right)\right). \tag{6.2}$$

The number of training iterations is 2000, the initial midpoint of weight vector is selected randomly from $[-0.01, 0.01]$, and two neurons are selected in the hidden layer. The fix learning rate is $\eta = 0.2$, $\beta = 0.5$, and $\mu = 0.5$.

In the learning procedure for the interval neural network, we clearly see from Figure 1(a) that the gradient norm is not convergent. Figure 2(a) shows that the error function $D$ is oscillating and not convergent. On the contrary, we see from Figure 1(b) that the gradient norm of the smoothing interval neural network is convergent. Figure 2(b) shows that the error function $D$, as well as $E$, is monotone decreasing and convergent.

From this numerical experiment, we can see that the proposed smoothing neural network can efficiently avoid the oscillation during the training process.

## Appendix

First, we give Lemmas A.1 and A.2. Then, we use them to prove Theorem 5.1.

**Lemma A.1.** *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*

$$S = [\gamma_1, \gamma_2]. \tag{A.1}$$

*Proof.* It is obvious that $\gamma_1 \leq \gamma_2$ and $S \subset [\gamma_1, \gamma_2]$. If $\gamma_1 = \gamma_2$, then (A.1) 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 $S \supset [\gamma_1, \gamma_2]$.

For any $a \in (\gamma_1, \gamma_2)$, there exists $\varepsilon > 0$ such that $(a - \varepsilon, a + \varepsilon) \subset (\gamma_1, \gamma_2)$. Noting $\lim_{m\to\infty}(b_{m+1} - b_m) = 0$, we observe that $b_m$ travels between $\gamma_1$ and $\gamma_2$ 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 $(\gamma_1, \gamma_2) \subset S$. Furthermore, $(\gamma_1, \gamma_2) \subset S$ immediately leads to $[\gamma_1, \gamma_2] \subset S$. This completes the proof. □

For any $k = 0, 1, 2, \ldots$, $1 \leq j \leq J$, we define the following notations.

$$\Phi^C_{0,k,j} = W^k_0 \cdot h^C_{k,j}, \qquad \Phi^R_{0,k,j} = \varphi\left(W^k_0\right) \cdot h^R_{k,j}, \qquad \Psi^C_{k,j} = h^C_{k+1,j} - h^C_{k,j}, \qquad \Psi^R_{k,j} = h^R_{k+1,j} - h^R_{k,j}. \tag{A.2}$$

**Lemma A.2.** *Suppose Assumption* (*A*2), (*A*3) *holds, for any* $k = 0, 1, 2, \ldots$ *and* $1 \leq j \leq J$, *then we have*

$$\max\left\{\left\|x^C_j\right\|, \left\|x^R_j\right\|, \left\|o^C_j\right\|, \left\|o^R_j\right\|\left\|W^k_0\right\|, \left\|\Phi^C_{0,k,j}\right\|, \left\|\Phi^R_{0,k,j}\right\|\right\} \leq M_0, \tag{A.3}$$

$$\sum_{j=1}^{J}\left(\beta f'^C_j\left(\Phi^C_{0,k,j}\right)h^C_{k,j}\Delta W^k_0 + (1-\beta)f'^R_j\left(\Phi^R_{0,k,j}\right)h^R_{k,j}\varphi'\left(W^k_0\right)\Delta W^k_0\right) = -\eta\left\|\frac{\partial E(W^k)}{\partial W_0}\right\|^2, \tag{A.4}$$

$$\sum_{j=1}^{J}\beta f'^C_j\left(\Phi^C_{0,k,j}\right)\left(\Delta W^k_0 \cdot \Psi^C_{k,j}\right) \leq M_1\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2, \tag{A.5}$$

$$\begin{aligned} &\sum_{j=1}^{J}\beta f'^C_j\left(\Phi^C_{0,k,j}\right)\left(W^k_0 \cdot \Psi^C_{k,j}\right) + \sum_{j=1}^{J}(1-\beta)f'^R_j\left(\phi^R_{0,k,j}\right)\left(\varphi\left(W^k_0\right) \cdot \Psi^R_{k,j}\right) \\ &\quad \leq \left(-\eta + M_2\eta^2\right)\sum_{m=1}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2, \end{aligned} \tag{A.6}$$

$$\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{\prime\prime C}\left(\xi_{0,k,j}^{C}\right)\left(\Phi_{0,k+1,j}^{C}-\Phi_{0,k,j}^{C}\right)^2 \le M_3\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2, \tag{A.7}$$

$$\sum_{j=1}^{J}(1-\beta) f_j^{\prime R}\left(\phi_{0,k,j}^{R}\right)\varphi'\left(\zeta_1^k\right)\left(\Delta W_0^k\cdot\Psi_{k,j}^{R}\right)\le M_4\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2, \tag{A.8}$$

$$\frac{1}{2}\sum_{j=1}^{J}(1-\beta) f_j^{\prime R}\left(\phi_{0,k,j}^{R}\right)\varphi''\left(\zeta_2^k\right)\left(\left(\Delta W_0^k\right)^2\cdot h_{k,j}^{R}\right)\le M_5\eta^2\left\|\frac{\partial E(W^k)}{\partial W_0}\right\|^2, \tag{A.9}$$

$$\frac{1}{2}\sum_{j=1}^{J}(1-\beta) f_j^{\prime\prime R}\left(\xi_{0,k,j}^{R}\right)\left(\Phi_{0,k+1,j}^{R}-\Phi_{0,k,j}^{R}\right)^2 \le M_6\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2, \tag{A.10}$$

*where* $M_i$ $(i = 0,1,2,3,4,5,6)$ *is independent of* $k$ *and* $j$, $\xi_{0,k,j}^{C}$ *lies on the segment between* $\Phi_{0,k+1,j}^{C}$ *and* $\Phi_{0,k,j}^{C}$, $\xi_{0,k,j}^{R}$ *lies on the segment between* $\Phi_{0,k+1,j}^{R}$ *and* $\Phi_{0,k,j}^{R}$, $\zeta_1^k, \zeta_2^k$ *both lie on the segment between* $W_0^{k+1}$ *and* $W_0^k$.

*Proof.* The proof of (A.3) in Lemma A.2: For the given training sample set, by Assumption $(A2)$, (4.2), and (4.4), it is easy to known that (A.3) is valid.

The proof of (A.4) in Lemma A.2: by (4.9) and (4.14), we have

$$\begin{aligned}&\sum_{j=1}^{J}\left(\beta f_j^{\prime C}\left(\Phi_{0,k,j}^{C}\right)h_{k,j}^{C}\Delta W_0^k+(1-\beta) f_j^{\prime R}\left(\Phi_{0,k,j}^{R}\right)h_{k,j}^{R}\varphi'\left(W_0^k\right)\Delta W_0^k\right)\\&\quad=\frac{\partial E(W^k)}{\partial W_0}\cdot\left(-\eta\frac{\partial E(W^k)}{\partial W_0}\right)=-\eta\left\|\frac{\partial E(W^k)}{\partial W_0}\right\|^2.\end{aligned}\tag{A.11}$$

This proves (A.4).

The proof of (A.5) in Lemma A.2: using the Mean Value Theorem, for any $1 \le m \le M$, $1 \le j \le J$, and $k = 0,1,2,\ldots$, we have

$$\begin{aligned}\Psi_{k,j,m}^{C}&=h_{k+1,j,m}^{C}-h_{k,j,m}^{C}\\&=\frac{1}{2}\left(f\left(s_{k+1,j,m}^{C}-s_{k+1,j,m}^{R}\right)-f\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)+f\left(s_{k+1,j,m}^{C}+s_{k+1,j,m}^{R}\right)-f\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\right)\\&=\frac{1}{2}\left(f'\left(t_{k,j,m}^{1}\right)\left(\left(s_{k+1,j,m}^{C}-s_{k+1,j,m}^{R}\right)-\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\right)\right.\\&\qquad\left.+f'\left(t_{k,j,m}^{2}\right)\left(\left(s_{k+1,j,m}^{C}+s_{k+1,j,m}^{R}\right)-\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\right)\right),\end{aligned}\tag{A.12}$$

where $t^1_{k,j,m}$ is on the segment between $s^C_{k+1,j,m} - s^R_{k+1,j,m}$ and $s^C_{k,j,m} - s^R_{k,j,m}$, $t^2_{k,j,m}$ is on the segment between $s^C_{k+1,j,m} + s^R_{k+1,j,m}$ and $s^C_{k,j,m} + s^R_{k,j,m}$. By (A.3), we have

$$\begin{aligned}\left|\Psi^C_{k,j,m}\right| &\le \frac{M_0}{2}\left(\left|\left(s^C_{k+1,j,m} - s^R_{k+1,j,m}\right) - \left(s^C_{k,j,m} - s^R_{k,j,m}\right)\right| + \left|\left(s^C_{k+1,j,m} + s^R_{k+1,j,m}\right) - \left(s^C_{k,j,m} + s^R_{k,j,m}\right)\right|\right) \\ &\le \frac{M_0}{2}\left(\left|s^C_{k+1,j,m} - s^C_{k,j,m}\right| + \left|s^R_{k+1,j,m} - s^R_{k,j,m}\right| + \left|s^C_{k+1,j,m} - s^C_{k,j,m}\right| + \left|s^R_{k+1,j,m} - s^R_{k,j,m}\right|\right) \\ &= M_0\left(\left|s^C_{k+1,j,m} - s^C_{k,j,m}\right| + \left|s^R_{k+1,j,m} - s^R_{k,j,m}\right|\right) \\ &= M_0\left(\left|\Delta W^k_m x^C_j\right| + \left|\left(\varphi\left(W^{k+1}_m\right) - \varphi\left(W^k_m\right)\right)x^R_j\right|\right) \\ &\le M_0^2\left(\left\|\Delta W^k_m\right\| + \left\|\varphi'\left(\tau^k_{1,m}\right)\right\|\left\|\Delta W^k_m\right\|\right),\end{aligned} \tag{A.13}$$

where $\tau^k_{1,m}$ is on the segment between $W^{k+1}_m$ and $W^k_m$. Since

$$\varphi(x) = \begin{cases} -x, & \text{if } x \le -\mu, \\ \widehat{\varphi}(x), & \text{if } -\mu < x < \mu, \\ x, & \text{if } x \ge \mu, \end{cases} \tag{A.14}$$

if $x \le -\mu$ and $x \ge \mu$, $|\varphi'(x)| = 1$, $|\varphi''(x)| = 0$.

If $-\mu < x < \mu$, we have

$$\begin{aligned}\varphi'(x) &= -\frac{1}{2\mu^3}x^3 + \frac{3}{2\mu}x \in (-1, 1), \\ \varphi''(x) &= -\frac{3}{2\mu^3}x^2 + \frac{3}{2\mu} \in \left(0, \frac{3}{2\mu}\right),\end{aligned} \tag{A.15}$$

so if $x \in \mathbb{R}$, we have

$$\left|\varphi'(x)\right| \le 1, \qquad \left|\varphi''(x)\right| \le \frac{3}{2\mu}. \tag{A.16}$$

According to (A.16) and (A.13), we can obtain that

$$\left|\Psi^C_{k,j,m}\right| \le 2M_0^2\left\|\Delta W^k_m\right\|. \tag{A.17}$$

By (A.17), for any $1 \leq j \leq J$ and $k = 0, 1, 2, \ldots$, we have

$$\left\| \Psi_{k,j}^{C} \right\|^2 = \left\| \begin{pmatrix} h_{k+1,j,1}^{C} - h_{k,j,1}^{C} \\ h_{k+1,j,2}^{C} - h_{k,j,2}^{C} \\ \vdots \\ h_{k+1,j,M}^{C} - h_{k,j,M}^{C} \end{pmatrix} \right\|^2 \leq 4M_0^4 \sum_{m=1}^{M} \left\| \Delta W_m^k \right\|^2. \tag{A.18}$$

According to the definition of $f_j^C(t)$, we get that $f_j^{'C}(t) = t_j^C - o_j^C$, combining with (A.3), we deduce that $|f_j^{'C}(\Phi_{0,k,j}^C)| \leq 2M_0$. By (A.18), we have

$$\begin{aligned} \sum_{j=1}^{J} \beta f_j^{'C} \left( \Phi_{0,k,j}^{C} \right) \left( \Delta W_0^k \cdot \Psi_{k,j}^{C} \right) &\leq 2\beta M_0 \sum_{j=1}^{J} \left\| \Delta W_0^k \right\| \left\| \Psi_{k,j}^{C} \right\| \\ &\leq \beta M_0 \sum_{j=1}^{J} \left( \left\| \Delta W_0^k \right\|^2 + \left\| \Psi_{k,j}^{C} \right\|^2 \right) \\ &\leq \beta J M_0 \left\| \Delta W_0^k \right\|^2 + 4\beta J M_0^5 \sum_{m=1}^{M} \left\| \Delta W_m^k \right\|^2 \\ &\leq M_1 \sum_{m=0}^{M} \left\| \Delta W_m^k \right\|^2 \\ &= M_1 \eta^2 \left\| \frac{\partial E(W^k)}{\partial W} \right\|^2, \end{aligned} \tag{A.19}$$

where $M_1 = \beta J M_0 \max\{1, 4M_0^4\}$. This proves (A.5).

The proof of (A.6) in Lemma A.2: using the Taylor expansion, we get that

$$\begin{aligned} \Psi_{k,j,m}^{C} &= h_{k+1,j,m}^{C} - h_{k,j,m}^{C} \\ &= \frac{1}{2} \left( f\left( s_{k+1,j,m}^{C} - s_{k+1,j,m}^{R} \right) - f\left( s_{k,j,m}^{C} - s_{k,j,m}^{R} \right) + f\left( s_{k+1,j,m}^{C} + s_{k+1,j,m}^{R} \right) - f\left( s_{k,j,m}^{C} + s_{k,j,m}^{R} \right) \right) \\ &= \frac{1}{2} \left( f'\left( s_{k,j,m}^{C} - s_{k,j,m}^{R} \right) \left( \left( s_{k+1,j,m}^{C} - s_{k+1,j,m}^{R} \right) - \left( s_{k,j,m}^{C} - s_{k,j,m}^{R} \right) \right) \right. \\ &\quad + f''\left( t_{k,j,m}^{3} \right) \left( \left( s_{k+1,j,m}^{C} - s_{k+1,j,m}^{R} \right) - \left( s_{k,j,m}^{C} - s_{k,j,m}^{R} \right) \right)^2 + f'\left( s_{k,j,m}^{C} + s_{k,j,m}^{R} \right) \\ &\quad \times \left( \left( s_{k+1,j,m}^{C} + s_{k+1,j,m}^{R} \right) - \left( s_{k,j,m}^{C} + s_{k,j,m}^{R} \right) \right) \\ &\quad \left. + f''\left( t_{k,j,m}^{4} \right) \left( \left( s_{k+1,j,m}^{C} + s_{k+1,j,m}^{R} \right) - \left( s_{k,j,m}^{C} + s_{k,j,m}^{R} \right) \right)^2 \right), \end{aligned} \tag{A.20}$$

where $t^3_{k,j,m}$ is on the segment between $s^C_{k+1,j,m}-s^R_{k+1,j,m}$ and $s^C_{k,j,m}-s^R_{k,j,m}$, $t^4_{k,j,m}$ is on the segment between $s^C_{k+1,j,m}+s^R_{k+1,j,m}$ and $s^C_{k,j,m}+s^R_{k,j,m}$. By (A.3), (A.16), we deduce that

$$\begin{aligned}
&\left(s^C_{k+1,j,m}-s^R_{k+1,j,m}\right)-\left(s^C_{k,j,m}-s^R_{k,j,m}\right)\\
&\quad=\Delta W^k_m x^C_j-\left(\phi'\left(W^k_m\right)\Delta W^k_m+\phi''\left(\tau^k_{2,m}\right)\left(\Delta W^k_m\right)^2\right)x^R_j\\
&\quad=\left(x^C_j-\phi'\left(W^k_m\right)x^R_j\right)\Delta W^k_m-\phi''\left(\tau^k_{2,m}\right)\left(\Delta W^k_m\right)^2x^R_j,\\
&\left(\left(s^C_{k+1,j,m}-s^R_{k+1,j,m}\right)-\left(s^C_{k,j,m}-s^R_{k,j,m}\right)\right)^2\\
&\quad=\left(\Delta W^k_m x^C_j-\phi'\left(\tau^k_{3,m}\right)\Delta W^k_m x^R_j\right)^2=\left(\left(x^C_j-\phi'\left(\tau^k_{3,m}\right)x^R_j\right)\Delta W^k_m\right)^2,
\end{aligned}\tag{A.21}$$

where $\tau^k_{2,m}, \tau^k_{3,m}$ both lie on the segment between $W^{k+1}_m$ and $W^k_m$. Similarly, we can deduce that

$$\begin{gathered}
\left(s^C_{k+1,j,m}+s^R_{k+1,j,m}\right)-\left(s^C_{k,j,m}+s^R_{k,j,m}\right)=\left(x^C_j+\phi'\left(W^k_m\right)x^R_j\right)\Delta W^k_m+\phi''\left(\tau^k_{4,m}\right)\left(\Delta W^k_m\right)^2x^R_j,\\
\left(\left(s^C_{k+1,j,m}+s^R_{k+1,j,m}\right)-\left(s^C_{k,j,m}+s^R_{k,j,m}\right)\right)^2=\left(\left(x^C_j+\phi'\left(\tau^k_{5,m}\right)x^R_j\right)\Delta W^k_m\right)^2,
\end{gathered}\tag{A.22}$$

where $\tau^k_{4,m}, \tau^k_{5,m}$ both lie on the segment between $W^{k+1}_m$ and $W^k_m$. Combining with (A.20), we have

$$\begin{aligned}
\Psi^C_{k,j,m}&=h^C_{k+1,j,m}-h^C_{k,j,m}\\
&=\frac{1}{2}\Bigg(f'\left(s^C_{k,j,m}-s^R_{k,j,m}\right)\left(\left(x^C_j-\phi'\left(W^k_m\right)x^R_j\right)\Delta W^k_m-\phi''\left(\tau^k_{2,m}\right)\left(\Delta W^k_m\right)^2x^R_j\right)+f''\left(t^3_{k,j,m}\right)\\
&\qquad\times\left(\left(x^C_j-\phi'\left(\tau^k_{3,m}\right)x^R_j\right)\Delta W^k_m\right)^2+f'\left(s^C_{k,j,m}+s^R_{k,j,m}\right)\\
&\qquad\times\left(\left(x^C_j+\phi'\left(W^k_m\right)x^R_j\right)\Delta W^k_m+\phi''\left(\tau^k_{4,m}\right)\left(\Delta W^k_m\right)^2x^R_j\right)\\
&\qquad+f''\left(t^4_{k,j,m}\right)\left(\left(x^C_j+\phi'\left(\tau^k_{5,m}\right)x^R_j\right)\Delta W^k_m\right)^2\Bigg)\\
&=\frac{1}{2}\Bigg(\left(f'\left(s^C_{k,j,m}-s^R_{k,j,m}\right)\left(x^C_j-\phi'\left(W^k_m\right)x^R_j\right)+f'\left(s^C_{k,j,m}+s^R_{k,j,m}\right)\left(x^C_j+\phi'\left(W^k_m\right)x^R_j\right)\right)\Delta W^k_m\\
&\qquad-f'\left(s^C_{k,j,m}-s^R_{k,j,m}\right)\phi''\left(\tau^k_{2,m}\right)\left(\Delta W^k_m\right)^2x^R_j+f''\left(t^3_{k,j,m}\right)\left(\left(x^C_j-\phi'\left(\tau^k_{3,m}\right)x^R_j\right)\Delta W^k_m\right)^2\\
&\qquad+f'\left(s^C_{k,j,m}+s^R_{k,j,m}\right)\phi''\left(\tau^k_{4,m}\right)\left(\Delta W^k_m\right)^2x^R_j+f''\left(t^4_{k,j,m}\right)\left(\left(x^C_j+\phi'\left(\tau^k_{5,m}\right)x^R_j\right)\Delta W^k_m\right)^2\Bigg).
\end{aligned}\tag{A.23}$$

By (A.23), we get that

$$\begin{aligned}
W_0^k \cdot \Psi_{k,j}^C = \frac{1}{2}\sum_{m=1}^{M} w_{0,m}^k \Bigg( & \Big( f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\left(x_j^C - \phi'\left(W_m^k\right)x_j^R\right) \\
& + f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\left(x_j^C + \phi'\left(W_m^k\right)x_j^R\right)\Big)\Delta W_m^k \\
& - f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\phi''\left(\tau_{2,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
& + f''\left(t_{k,j,m}^3\right)\left(\left(x_j^C - \phi'\left(\tau_{3,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2 \\
& + f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\phi''\left(\tau_{4,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
& + f''\left(t_{k,j,m}^4\right)\left(\left(x_j^C + \phi'\left(\tau_{5,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2\Bigg) \\
= \Delta_1 + \Delta_2, &
\end{aligned} \tag{A.24}$$

where

$$\begin{aligned}
\Delta_1 = \frac{1}{2}\sum_{m=1}^{M} w_{0,m}^k \Big( & f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\left(x_j^C - \phi'\left(W_m^k\right)x_j^R\right) \\
& + f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\left(x_j^C + \phi'\left(W_m^k\right)x_j^R\right)\Big)\Delta W_m^k,
\end{aligned} \tag{A.25}$$

$$\begin{aligned}
\Delta_2 = \frac{1}{2}\sum_{m=1}^{M} w_{0,m}^k \Big( & -f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\phi''\left(\tau_{2,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
& + f''\left(t_{k,j,m}^3\right)\left(\left(x_j^C - \phi'\left(\tau_{3,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2 \\
& + f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\phi''\left(\tau_{4,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
& + f''\left(t_{k,j,m}^4\right)\left(\left(x_j^C + \phi'\left(\tau_{5,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2\Big).
\end{aligned} \tag{A.26}$$

This together with (A.25) leads to

$$\begin{aligned}
& \sum_{j=1}^{J} \beta f_j^{'C}\left(\Phi_{0,k,j}^C\right)\Delta_1 \\
& \quad = \frac{1}{2}\sum_{j=1}^{J} \beta f_j^{'C}\left(\Phi_{0,k,j}^C\right)\sum_{m=1}^{M} w_{0,m}^k \Big( f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\left(x_j^C - \phi'\left(W_m^k\right)x_j^R\right) \\
& \qquad + f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\left(x_j^C + \phi'\left(W_m^k\right)x_j^R\right)\Big)\Delta W_m^k.
\end{aligned} \tag{A.27}$$

This together with (A.26) leads to

$$\begin{aligned}
&\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\Delta_2\\
&\quad=\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^{k}\Bigg(-f'\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\phi''\left(\tau_{2,m}^{k}\right)\left(\Delta W_m^k\right)^2 x_j^R\\
&\qquad+f''\left(t_{k,j,m}^{3}\right)\left(\left(x_j^C-\phi'\left(\tau_{3,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\\
&\qquad+f'\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\phi''\left(\tau_{4,m}^{k}\right)\left(\Delta W_m^k\right)^2 x_j^R\\
&\qquad+f''\left(t_{k,j,m}^{4}\right)\left(\left(x_j^C+\phi'\left(\tau_{5,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\Bigg).
\end{aligned}\tag{A.28}$$

By (A.3), (A.16) and $|f_j^{'C}(\Phi_{0,k,j}^{C})|\le 2M_0$, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^{k}\left(-f'\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\phi''\left(\tau_{2,m}^{k}\right)\left(\Delta W_m^k\right)^2 x_j^R\right)\\
&\quad\le\frac{1}{2}\beta\sum_{j=1}^{J}\sum_{m=1}^{M}\left\|f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\right\|\cdot\left\|w_{0,m}^{k}\right\|\cdot\left\|f'\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\right\|\cdot\left\|\phi''\left(\tau_{2,m}^{k}\right)\right\|\cdot\left\|\Delta W_m^k\right\|^2\cdot\left\|x_j^R\right\|\\
&\quad\le\frac{1}{2}\beta\sum_{j=1}^{J}\sum_{m=1}^{M}2M_0\cdot M_0\cdot M_0\cdot\frac{3}{2\mu}\cdot M_0\cdot\left\|\Delta W_m^k\right\|^2\\
&\quad=\frac{3}{2\mu}\beta J M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2.
\end{aligned}\tag{A.29}$$

Similarly, we can obtain that

$$\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^{k}f''\left(t_{k,j,m}^{3}\right)\left(\left(x_j^C-\phi'\left(\tau_{3,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\\
&\quad\le 4\beta J M_0^5\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2,\\
&\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^{k}f'\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\phi''\left(\tau_{4,m}^{k}\right)\left(\Delta W_m^k\right)^2 x_j^R\\
&\quad\le\frac{3}{2\mu}\beta J M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2,
\end{aligned}$$

$$\begin{aligned}&\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^{k}f''\left(t_{k,j,m}^{4}\right)\left(\left(x_j^C+\phi'\left(\tau_{5,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\\&\quad\le 4\beta J M_0^5\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2.\end{aligned}\tag{A.30}$$

So by (A.28), (A.29), and (A.30), we have

$$\begin{aligned}&\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\Delta_2\\&\quad\le\frac{3}{2\mu}\beta J M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2+4\beta J M_0^5\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\\&\qquad+\frac{3}{2\mu}\beta J M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2+4\beta J M_0^5\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\\&\quad=\left(\frac{3}{\mu}+8M_0\right)\beta J M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2,\end{aligned}\tag{A.31}$$

with (A.23), similarly, we get that

$$\begin{aligned}\Psi_{k,j,m}^{R}&=h_{k+1,j,m}^{R}-h_{k,j,m}^{R}\\&=\frac{1}{2}\bigg(f'\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\Big(\left(x_j^C+\phi'\left(W_m^k\right)x_j^R\right)\Delta W_m^k+\phi''\left(\tau_{6,m}^{k}\right)\left(\Delta W_m^k\right)^2x_j^R\Big)+f''\left(t_{k,j,m}^{5}\right)\\&\qquad\times\left(\left(x_j^C+\phi'\left(\tau_{7,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2-f'\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\\&\qquad\times\Big(\left(x_j^C-\phi'\left(W_m^k\right)x_j^R\right)\Delta W_m^k-\phi''\left(\tau_{8,m}^{k}\right)\left(\Delta W_m^k\right)^2x_j^R\Big)\\&\qquad-f''\left(t_{k,j,m}^{6}\right)\left(\left(x_j^c-\phi'\left(\tau_{9,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\bigg)\\&=\frac{1}{2}\bigg(\Big(f'\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\left(x_j^C+\phi'\left(W_m^k\right)x_j^R\right)-f'\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\left(x_j^C-\phi'\left(W_m^k\right)x_j^R\right)\Big)\Delta W_m^k\\&\qquad+f'\left(s_{k,j,m}^{C}+s_{k,j,m}^{R}\right)\phi''\left(\tau_{6,m}^{k}\right)\left(\Delta W_m^k\right)^2x_j^R+f''\left(t_{k,j,m}^{5}\right)\left(\left(x_j^C+\phi'\left(\tau_{7,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\\&\qquad+f'\left(s_{k,j,m}^{C}-s_{k,j,m}^{R}\right)\phi''\left(\tau_{8,m}^{k}\right)\left(\Delta W_m^k\right)^2x_j^R\\&\qquad-f''\left(t_{k,j,m}^{6}\right)\left(\left(x_j^C-\phi'\left(\tau_{9,m}^{k}\right)x_j^R\right)\Delta W_m^k\right)^2\bigg),\end{aligned}\tag{A.32}$$

where $\tau_{6,m}^k, \tau_{7,m}^k, \tau_{8,m}^k, \tau_{9,m}^k$ lie on the segment between $W_m^{k+1}$ and $W_m^k$, $t_{k,j,m}^5$ lies on the segment between $s_{k+1,j,m}^C + s_{k+1,j,m}^R$ and $s_{k,j,m}^C + s_{k,j,m}^R$, $t_{k,j,m}^6$ lies on the segment between $s_{k+1,j,m}^C - s_{k+1,j,m}^R$ and $s_{k,j,m}^C - s_{k,j,m}^R$. By (A.32), we have

$$\begin{aligned}
&\varphi\left(W_0^k\right) \cdot \Psi_{k,j}^R \\
&\quad = \frac{1}{2}\sum_{m=1}^{M}\varphi\left(w_{0,m}^k\right)\Bigg(\bigg(f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\left(x_j^C + \phi'\left(W_m^k\right)x_j^R\right) \\
&\qquad\qquad -f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\left(x_j^C - \phi'\left(W_m^k\right)x_j^R\right)\bigg)\Delta W_m^k \\
&\qquad\qquad + f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\phi''\left(\tau_{6,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
&\qquad\qquad + f''\left(t_{k,j,m}^5\right)\left(\left(x_j^C + \phi'\left(\tau_{7,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2 \\
&\qquad\qquad + f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\phi''\left(\tau_{8,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
&\qquad\qquad -f''\left(t_{k,j,m}^6\right)\left(\left(x_j^C - \phi'\left(\tau_{9,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2\Bigg) \\
&\quad = \Delta_3 + \Delta_4,
\end{aligned} \tag{A.33}$$

where

$$\begin{aligned}
\Delta_3 = &\frac{1}{2}\sum_{m=1}^{M}\varphi\left(w_{0,m}^k\right) \\
&\times\left(\left(f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\left(x_j^C + \phi'\left(W_m^k\right)x_j^R\right)\right.\right. \\
&\qquad\left.\left. -f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\left(x_j^C - \phi'\left(W_m^k\right)x_j^R\right)\right)\Delta W_m^k\right),
\end{aligned} \tag{A.34}$$

$$\begin{aligned}
\Delta_4 = \frac{1}{2}\sum_{m=1}^{M}\varphi\left(w_{0,m}^k\right)\Bigg(&f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\phi''\left(\tau_{6,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
&+ f''\left(t_{k,j,m}^5\right)\left(\left(x_j^C + \phi'\left(\tau_{7,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2 \\
&+ f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\phi''\left(\tau_{8,m}^k\right)\left(\Delta W_m^k\right)^2 x_j^R \\
&-f''\left(t_{k,j,m}^6\right)\left(\left(x_j^C - \phi'\left(\tau_{9,m}^k\right)x_j^R\right)\Delta W_m^k\right)^2\Bigg).
\end{aligned} \tag{A.35}$$

By (A.34), we have

$$\begin{aligned}
&\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^R\right)\Delta_3 \\
&\quad = \frac{1}{2}\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^R\right)\sum_{m=1}^{M}\varphi\left(w_{0,m}^k\right)\left(\left(f'\left(s_{k,j,m}^C + s_{k,j,m}^R\right)\left(x_j^C + \phi'\left(W_m^k\right)x_j^R\right)\right.\right. \\
&\qquad\qquad\left.\left. -f'\left(s_{k,j,m}^C - s_{k,j,m}^R\right)\left(x_j^C - \phi'\left(W_m^k\right)x_j^R\right)\right)\Delta W_m^k\right),
\end{aligned} \tag{A.36}$$

with (A.31), similarly, this together with (A.35) leads to

$$\begin{aligned}
&\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^{R}\right)\Delta_4\\
&\quad\le \frac{3}{2\mu}(1-\beta)JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2+4(1-\beta)JM_0^5\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\\
&\qquad+\frac{3}{2\mu}(1-\beta)JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2+4(1-\beta)JM_0^5\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\\
&\quad=\left(\frac{3}{\mu}+8M_0\right)(1-\beta)JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2.
\end{aligned}\tag{A.37}$$

By (A.27), (A.31), (A.36) and (A.37), we obtain that

$$\begin{aligned}
&\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\left(W_0^k\cdot\Psi_{k,j}^{C}\right)+\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^{R}\right)\left(\varphi\left(W_0^k\right)\cdot\Psi_{k,j}^{R}\right)\\
&\quad=\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)(\Delta_1+\Delta_2)+\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^{R}\right)(\Delta_3+\Delta_4)\\
&\quad\le\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^k\Bigl(f'\left(s_{k,j,m}^C-s_{k,j,m}^R\right)\left(x_j^C-\phi'\left(W_m^k\right)x_j^R\right)\\
&\qquad+f'\left(s_{k,j,m}^C+s_{k,j,m}^R\right)\left(x_j^C+\phi'\left(W_m^k\right)x_j^R\right)\Bigr)\Delta W_m^k+\frac{1}{2}\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^{R}\right)\sum_{m=1}^{M}\varphi\left(w_{0,m}^k\right)\\
&\qquad\times\Bigl(\Bigl(f'\left(s_{k,j,m}^C+s_{k,j,m}^R\right)\left(x_j^C+\phi'\left(W_m^k\right)x_j^R\right)-f'\left(s_{k,j,m}^C-s_{k,j,m}^R\right)\left(x_j^C-\phi'\left(W_m^k\right)x_j^R\right)\Bigr)\Delta W_m^k\Bigr)\\
&\qquad+\left(\frac{3}{\mu}+8M_0\right)\beta JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2+\left(\frac{3}{\mu}+8M_0\right)(1-\beta)JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\\
&\quad=\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{'C}\left(\Phi_{0,k,j}^{C}\right)\sum_{m=1}^{M}w_{0,m}^k\Bigl(f'\left(s_{k,j,m}^C-s_{k,j,m}^R\right)\left(x_j^C-\phi'\left(W_m^k\right)x_j^R\right)\\
&\qquad\qquad\qquad\qquad+f'\left(s_{k,j,m}^C+s_{k,j,m}^R\right)\left(x_j^C+\phi'\left(W_m^k\right)x_j^R\right)\Bigr)\\
&\qquad\times\Delta W_m^k+\frac{1}{2}\sum_{j=1}^{J}(1-\beta)f_j^{'R}\left(\phi_{0,k,j}^{R}\right)\sum_{m=1}^{M}\varphi\left(w_{0,m}^k\right)\\
&\qquad\times\Bigl(f'\left(s_{k,j,m}^C+s_{k,j,m}^R\right)\left(x_j^C+\phi'\left(W_m^k\right)x_j^R\right)-f'\left(s_{k,j,m}^C-s_{k,j,m}^R\right)\left(x_j^C-\phi'\left(W_m^k\right)x_j^R\right)\Bigr)\Delta W_m^k\\
&\qquad+\left(\frac{3}{\mu}+8M_0\right)JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2.
\end{aligned}\tag{A.38}$$

Combining with (4.11), (4.12), and (4.14), we get that

$$\begin{aligned}
&\sum_{j=1}^{J}\beta f_j^{\prime C}\left(\Phi_{0,k,j}^{C}\right)\left(W_0^k\cdot\Psi_{k,j}^{C}\right)+\sum_{j=1}^{J}(1-\beta)f_j^{\prime R}\left(\phi_{0,k,j}^{R}\right)\left(\varphi\left(W_0^k\right)\cdot\Psi_{k,j}^{R}\right)\\
&\quad\le\sum_{m=1}^{M}\frac{\partial E(W^k)}{\partial W_m}\cdot\Delta W_m^k+\left(\frac{3}{\mu}+8M_0\right)JM_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\\
&\quad=-\eta\sum_{m=1}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2+M_2\eta^2\sum_{m=1}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2\\
&\quad=\left(-\eta+M_2\eta^2\right)\sum_{m=1}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2,
\end{aligned}\tag{A.39}$$

where $M_2=((3/\mu)+8M_0)JM_0^4$. This proves (A.6).

The proof of (A.7) in Lemma A.2: According to the definition of $f_j^C(t)$, we get that $f_j^{\prime\prime C}(t)=1$, combining with (A.3), (A.18), we have

$$\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{J}\beta f_j^{\prime\prime C}\left(\xi_{0,k,j}^{C}\right)\left(\Phi_{0,k+1,j}^{C}-\Phi_{0,k,j}^{C}\right)^2\\
&\quad=\frac{1}{2}\beta\sum_{j=1}^{J}\left\|\Phi_{0,k+1,j}^{C}-\Phi_{0,k,j}^{C}\right\|^2\\
&\quad=\frac{1}{2}\beta\sum_{j=1}^{J}\left\|W_0^{k+1}\cdot h_{k+1,j}^{C}-W_0^k\cdot h_{k,j}^{C}\right\|^2\\
&\quad=\frac{1}{2}\beta\sum_{j=1}^{J}\left\|\left(W_0^{k+1}-W_0^k\right)\cdot h_{k+1,j}^{C}+W_0^k\cdot\left(h_{k+1,j}^{C}-h_{k,j}^{C}\right)\right\|^2\\
&\quad\le\beta\sum_{j=1}^{J}\left(M_0^2\left\|\Delta W_0^k\right\|^2+M_0^2\left\|\Psi_{k,j}^{C}\right\|^2\right)\\
&\quad\le\beta JM_0^2\left(\left\|\Delta W_0^k\right\|^2+4M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\right)\\
&\quad\le M_3\sum_{m=0}^{M}\left\|\Delta W_m^k\right\|^2\\
&\quad=M_3\eta^2\sum_{m=0}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2\\
&\quad=M_3\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2,
\end{aligned}\tag{A.40}$$

where $M_3=\beta JM_0^2\max\{1,4M_0^4\}$. This proves (A.7).

The proof of (A.8) in Lemma A.2: With (A.17), similarly, for any $1 \le j \le J$ and $k = 0, 1, 2, \ldots$, we can get that

$$\left\| \Psi^R_{k,j} \right\|^2 \le 4M_0^4 \sum_{m=1}^{M} \left\| \Delta W_m^k \right\|^2. \tag{A.41}$$

According to the definition of $f_j^R(t)$, we get that $f_j^{\prime R}(t) = t_j^R - o_j^R$, combining with (A.3), we can obtain that $|f_j^{\prime R}(\Phi^R_{0,k,j})| \le 2M_0$. By (A.16) and (A.41), we deduce that

$$\begin{aligned}
&\sum_{j=1}^{J} (1-\beta) f_j^{\prime R}\left(\Phi^R_{0,k,j}\right) \varphi'\left(\zeta_1^k\right)\left(\Delta W_0^k \cdot \Psi^R_{k,j}\right) \\
&\quad \le 2(1-\beta) M_0 \sum_{j=1}^{J} \left\| \Delta W_0^k \right\| \left\| \Psi^R_{k,j} \right\| \\
&\quad \le (1-\beta) M_0 \sum_{j=1}^{J} \left( \left\| \Delta W_0^k \right\|^2 + \left\| \Psi^R_{k,j} \right\|^2 \right) \\
&\quad \le (1-\beta) J M_0 \left\| \Delta W_0^k \right\|^2 + 4(1-\beta) J M_0^5 \sum_{m=1}^{M} \left\| \Delta W_m^k \right\|^2 \\
&\quad \le M_4 \sum_{m=0}^{M} \left\| \Delta W_m^k \right\|^2 \\
&\quad = M_4 \eta^2 \left\| \frac{\partial E(W^k)}{\partial W} \right\|^2,
\end{aligned} \tag{A.42}$$

where $M_4 = (1-\beta) J M_0 \max\{1, 4M_0^4\}$. This proves (A.8).

The proof of (A.9) in lemma A.2: By $|f_j^{\prime R}(\Phi^R_{0,k,j})| \le 2M_0$, (A.3) and (A.16), we get that

$$\begin{aligned}
&\frac{1}{2} \sum_{j=1}^{J} (1-\beta) f_j^{\prime R}\left(\phi^R_{0,k,j}\right) \varphi''\left(\zeta_2^k\right)\left(\left(\Delta W_0^k\right)^2 \cdot h^R_{k,j}\right) \\
&\quad \le (1-\beta) M_0 \cdot \frac{3}{2\mu} \sum_{j=1}^{J} \left\| \Delta W_0^k \right\|^2 \cdot \left\| h^R_{k,j} \right\| \\
&\quad \le \frac{3}{2\mu} (1-\beta) J M_0^2 \left\| \Delta W_0^k \right\|^2 \\
&\quad \le M_5 \eta^2 \left\| \frac{\partial E(W^k)}{\partial W_0} \right\|^2,
\end{aligned} \tag{A.43}$$

where $M_5 = (3/2\mu)(1-\beta) J M_0^2$. This proves (A.9).

The proof of (A.10) in Lemma A.2: According to the definition of $f_j^R(t)$, we get that $f_j^{\prime\prime R}(t) = 1$, combining with (A.3) and (A.41), we have

$$\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{J}(1-\beta)f_j^{\prime\prime R}\left(\xi_{0,k,j}^R\right)\left(\Phi_{0,k+1,j}^R - \Phi_{0,k,j}^R\right)^2 \\
&\quad = \frac{1}{2}(1-\beta)\sum_{j=1}^{J}\left\|\Phi_{0,k+1,j}^R - \Phi_{0,k,j}^R\right\|^2 \\
&\quad = \frac{1}{2}(1-\beta)\sum_{j=1}^{J}\left\|\varphi\left(W_0^{k+1}\right)\cdot h_{k+1,j}^R - \varphi\left(W_0^k\right)\cdot h_{k,j}^R\right\|^2 \\
&\quad = \frac{1}{2}(1-\beta)\sum_{j=1}^{J}\left\|\left(\varphi\left(W_0^{k+1}\right) - \varphi\left(W_0^k\right)\right)\cdot h_{k+1,j}^R + \varphi\left(W_0^k\right)\cdot\left(h_{k+1,j}^R - h_{k,j}^R\right)\right\|^2 \\
&\quad \leq (1-\beta)\sum_{j=1}^{J}\left(M_0^2\left\|\Delta W_0^k\right\|^2 + M_0^2\left\|\Psi_{k,j}^R\right\|^2\right) \qquad \text{(A.44)} \\
&\quad \leq (1-\beta)JM_0^2\left(\left\|\Delta W_0^k\right\|^2 + 4M_0^4\sum_{m=1}^{M}\left\|\Delta W_m^k\right\|^2\right) \\
&\quad \leq M_6\sum_{m=0}^{M}\left\|\Delta W_m^k\right\|^2 \\
&\quad = M_6\eta^2\sum_{m=0}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2 \\
&\quad = M_6\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2,
\end{aligned}$$

where $M_6 = (1-\beta)JM_0^2\max\{1, 4M_0^4\}$. This proves (A.10). Thus this completes the proof of Lemma A.2. □

Now we are ready to prove Theorem 5.1.

*Proof.* Using the Taylor expansion and Lemma A.2, for any $k = 0, 1, 2, \ldots$, we have

$$\begin{aligned}
&E\left(W^{k+1}\right) - E\left(W^k\right) \\
&\quad = \sum_{j=1}^{J}\left(\beta f_j^C\left(\Phi_{0,k+1,j}^C\right) + (1-\beta)f_j^R\left(\Phi_{0,k+1,j}^R\right) - \beta f_j^C\left(\Phi_{0,k,j}^C\right) - (1-\beta)f_j^R\left(\Phi_{0,k,j}^R\right)\right) \\
&\quad = \sum_{j=1}^{J}\left(\beta f_j^{\prime C}\left(\Phi_{0,k,j}^C\right)\left(\Phi_{0,k+1,j}^C - \Phi_{0,k,j}^C\right) + \frac{1}{2}\beta f_j^{\prime\prime C}\left(\xi_{0,k,j}^C\right)\left(\Phi_{0,k+1,j}^C - \Phi_{0,k,j}^C\right)^2\right. \\
&\qquad\left. +(1-\beta)f_j^{\prime R}\left(\Phi_{0,k,j}^R\right)\left(\Phi_{0,k+1,j}^R - \Phi_{0,k,j}^R\right) + \frac{1}{2}(1-\beta)f_j^{\prime\prime R}\left(\xi_{0,k,j}^R\right)\left(\Phi_{0,k+1,j}^R - \Phi_{0,k,j}^R\right)^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{J}\Bigg(\beta f_j^{\prime C}\left(\Phi_{0,k,j}^{C}\right)\left(W_0^{k+1}h_{k+1,j}^{C} - W_0^{k}h_{k,j}^{C}\right) + \frac{1}{2}\beta f_j^{\prime\prime C}\left(\xi_{0,k,j}^{C}\right)\left(\Phi_{0,k+1,j}^{C} - \Phi_{0,k,j}^{C}\right)^2 \\
&\qquad + (1-\beta) f_j^{\prime R}\left(\Phi_{0,k,j}^{R}\right)\left(\varphi\left(W_0^{k+1}\right)h_{k+1,j}^{R} - \varphi\left(W_0^{k}\right)h_{k,j}^{R}\right) \\
&\qquad + \frac{1}{2}(1-\beta) f_j^{\prime\prime R}\left(\xi_{0,k,j}^{R}\right)\left(\Phi_{0,k+1,j}^{R} - \Phi_{0,k,j}^{R}\right)^2\Bigg) \\
&= \sum_{j=1}^{J}\Bigg(\beta f_j^{\prime C}\left(\Phi_{0,k,j}^{C}\right)\left(\Delta W_0^{k}\cdot h_{k,j}^{C} + W_0^{k}\cdot \Psi_{k,j}^{C} + \Delta W_0^{k}\cdot \Psi_{k,j}^{C}\right) \\
&\qquad + \frac{1}{2}\beta f_j^{\prime\prime C}\left(\xi_{0,k,j}^{C}\right)\left(\Phi_{0,k+1,j}^{C} - \Phi_{0,k,j}^{C}\right)^2 + (1-\beta) f_j^{\prime R}\left(\Phi_{0,k,j}^{R}\right) \\
&\qquad \times \left(\left(\varphi\left(W_0^{k+1}\right) - \varphi\left(W_0^{k}\right)\right)\cdot h_{k,j}^{R} + \varphi\left(W_0^{k}\right)\cdot \Psi_{k,j}^{R} + \left(\varphi\left(W_0^{k+1}\right) - \varphi\left(W_0^{k}\right)\right)\cdot \Psi_{k,j}^{R}\right) \\
&\qquad + \frac{1}{2}(1-\beta) f_j^{\prime\prime R}\left(\xi_{0,k,j}^{R}\right)\left(\Phi_{0,k+1,j}^{R} - \Phi_{0,k,j}^{R}\right)^2\Bigg) \\
&= \sum_{j=1}^{J}\Bigg(\beta f_j^{\prime C}\left(\Phi_{0,k,j}^{C}\right)\left(\Delta W_0^{k}\cdot h_{k,j}^{C} + W_0^{k}\cdot \Psi_{k,j}^{C} + \Delta W_0^{k}\cdot \Psi_{k,j}^{C}\right) \\
&\qquad + \frac{1}{2}\beta f_j^{\prime\prime C}\left(\xi_{0,k,j}^{C}\right)\left(\Phi_{0,k+1,j}^{C} - \Phi_{0,k,j}^{C}\right)^2 + (1-\beta) f_j^{\prime R}\left(\Phi_{0,k,j}^{R}\right) \\
&\qquad \times \left(\left(\varphi'\left(W_0^{k}\right)\Delta W_0^{k} + \varphi''\left(\zeta_2^{k}\right)\left(\Delta W_0^{k}\right)^2\right)\cdot h_{k,j}^{R} + \varphi\left(W_0^{k}\right)\cdot \Psi_{k,j}^{R} + \varphi'\left(\zeta_1^{k}\right)\Delta W_0^{k}\cdot \Psi_{k,j}^{R}\right) \\
&\qquad + \frac{1}{2}(1-\beta) f_j^{\prime\prime R}\left(\xi_{0,k,j}^{R}\right)\left(\Phi_{0,k+1,j}^{R} - \Phi_{0,k,j}^{R}\right)^2\Bigg) \\
&\le -\eta\left\|\frac{\partial E(W^k)}{\partial W_0}\right\|^2 + \left(-\eta + M_2\eta^2\right)\sum_{m=1}^{M}\left\|\frac{\partial E(W^k)}{\partial W_m}\right\|^2 + M_1\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2 + M_4\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2 \\
&\quad + M_5\eta^2\left\|\frac{\partial E(W^k)}{\partial W_0}\right\|^2 + M_3\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2 + M_6\eta^2\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2 \\
&\le -\left(\eta - M_7\eta^2\right)\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2,
\end{aligned}
\tag{A.45}
$$

where $M_7 = M_1+M_2+M_3+M_4+M_5+M_6$, $\xi_{0,k,j}^{C}$ lies on the segment between $\Phi_{0,k+1,j}^{C}$ and $\Phi_{0,k,j}^{C}$, $\xi_{0,k,j}^{R}$ lies on the segment between $\Phi_{0,k+1,j}^{R}$ and $\Phi_{0,k,j}^{R}$, $\zeta_1^k$, $\zeta_2^k$ both lie on the segment between $W_0^{k+1}$ and $W_0^k$. Let $\gamma = \eta - M_7\eta^2$, then

$$
E\left(W^{k+1}\right) \le E\left(W^{k}\right) - \gamma\left\|\frac{\partial E(W^k)}{\partial W}\right\|^2. \tag{A.46}
$$

Obviously, we require the learning rate $\eta$ to satisfy

$$0 < \eta < \frac{1}{M_7}. \tag{A.47}$$

Thus, we can obtain that

$$E\left(W^{k+1}\right) \leq E\left(W^k\right), \quad k = 0, 1, 2, \ldots. \tag{A.48}$$

This together with (A.46) leads to

$$\begin{aligned} E\left(W^{k+1}\right) &\leq E\left(W^k\right) - \gamma \left\| \frac{\partial E(W^k)}{\partial W} \right\|^2 \\ &\leq \cdots \leq E\left(W^0\right) - \gamma \sum_{t=0}^{k} \left\| \frac{\partial E(W^t)}{\partial W} \right\|^2. \end{aligned} \tag{A.49}$$

Since $E(W^{k+1}) \geq 0$, we have

$$\gamma \sum_{t=0}^{k} \leq E\left(W^0\right). \tag{A.50}$$

Letting $k \rightarrow \infty$ results in

$$\sum_{t=0}^{\infty} \left\| \frac{\partial E(W^t)}{\partial W} \right\|^2 \leq \frac{1}{\gamma} E\left(W^0\right) < \infty. \tag{A.51}$$

So this immediately gives

$$\lim_{k \rightarrow \infty} \left\| \frac{\partial E(W^k)}{\partial W} \right\|^2 = 0. \tag{A.52}$$

According to (4.14) and (A.52), we get that

$$\lim_{k \rightarrow \infty} \left\| \Delta W^k \right\| = 0. \tag{A.53}$$

According to $(A1)$, 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 (5.3) and the continuity of $E_{\mathbf{w}}(\mathbf{w})$ that

$$\|E_{\mathbf{w}}(\mathbf{w}^*)\| = \lim_{k \rightarrow \infty} \|E_{\mathbf{w}}(\mathbf{w}^{m_k})\| = \lim_{m \rightarrow \infty} \|E_{\mathbf{w}}(\mathbf{w}^m)\| = 0. \tag{A.54}$$

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 the contrary that $\{\mathbf{w}^m\}$ has at least two accumulation points $\overline{\mathbf{w}} \neq \tilde{\mathbf{w}}$. We write $\mathbf{w}^m = (w_1^m, w_2^m, \ldots, w_{n(p+1)}^m)^T$. It is easy to see from (4.13) and (4.14) that $\lim_{m\to\infty}\|\mathbf{w}^{m+1} - \mathbf{w}^m\| = 0$, 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 $\tilde{\mathbf{w}}$ do not equal to each other, that is, $\overline{w}_1 \neq \tilde{w}_1$. For any real number $\lambda \in (0,1)$, let $w_1^\lambda = \lambda\overline{w}_1 + (1-\lambda)\tilde{w}_1$. By Lemma A.1, there exists a subsequence $\{w_1^{m_{k_1}}\}$ of $\{w_1^m\}$ converging to $w_1^\lambda$ as $k_1 \to \infty$. Due to the boundedness of $\{w_2^{m_{k_1}}\}$, there is a convergent subsequence $\{w_2^{m_{k_2}}\} \subset \{w_2^{m_{k_1}}\}$. 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 proves (5.4). Thus this completes the proof of Theorem 5.1. □

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (11171367) and the Fundamental Research Funds for the Central Universities of China.

## References

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[2] M. Perez, "Artificial neural networks and bankruptcy forecasting: a state of the art," *Neural Computing and Applications*, vol. 15, no. 2, pp. 154–163, 2006.

[3] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: the state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[4] M. W. Craven and J. W. Shavlik, "Using neural networks for data mining," *Future Generation Computer Systems*, vol. 13, no. 2-3, pp. 211–229, 1997.

[5] S. K. Pal, V. Talwar, and P. Mitra, "Web mining in soft computing framework: relevance, state of the art and future directions," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1163–1177, 2002.

[6] K. I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, 1989.

[7] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[8] H. White, "Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, vol. 3, no. 5, pp. 535–549, 1990.

[9] H. Ishibuchi and H. Tanaka, "An extension of the BP algorithm to interval input vectors," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'91)*, vol. 2, pp. 1588–1593, Singapore, 1991.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[11] D. E. Rumelhart, J. L. McClelland, and The PDP Research Group, *Parallel Distributed Processing*, vol. 1, MIT Press, Cambridge, Mass, USA, 1986.

[12] C. A. Hernandez, J. Espi, K. Nakayama, and M. Fernandez, "Interval arithmetic backpropagation," in *Proceedings of International Joint Conference on Neural Networks*, vol. 1, pp. 375–378, Nagoya, Japan, October 1993.

[13] A. M. S. Roque, C. Maté, J. Arroyo, and A. Sarabia, “IMLP: applying multi-layer perceptrons to interval-valued data,” *Neural Processing Letters*, vol. 25, no. 2, pp. 157–169, 2007.

[14] H. M. Shao and G. F. Zheng, “Convergence analysis of a back-propagation algorithm with adaptive momentum,” *Neurocomputing*, vol. 74, no. 5, pp. 749–752, 2011.

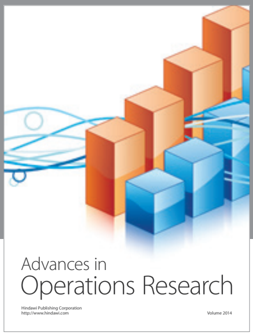
[15] W. Wu, J. Wang, M. S. Cheng, and Z. X. Li, “Convergence analysis of online gradient method for BP neural networks,” *Neural Networks*, vol. 24, no. 1, pp. 91–98, 2011.

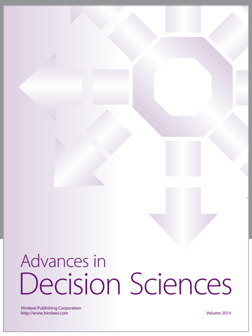
[16] D. P. Xu, H. S. Zhang, and L. J. Liu, “Convergence analysis of three classes of split-complex gradient algorithms for complex-valued recurrent neural networks,” *Neural Computation*, vol. 22, no. 10, pp. 2655–2677, 2010.

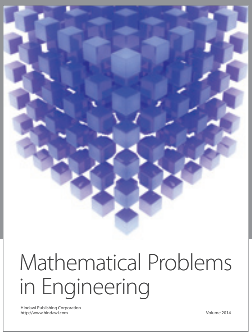
[17] J. Wang, J. Yang, and W. Wu, “Convergence of cyclic and almost-cyclic learning with momentum for feedforward neural networks,” *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1297–1306, 2011.

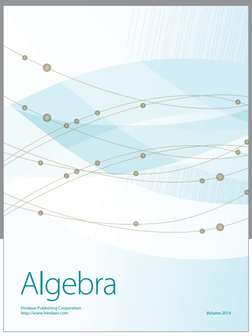
[18] R. E. Moore, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1966.

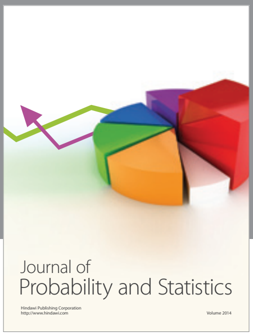
[19] T. Sunaga, “Theory of an interval algebra and its applications to numerical analysis,” *RAAG Memoirs*, vol. 2, pp. 29–46, 1958.


Advances in
Operations Research

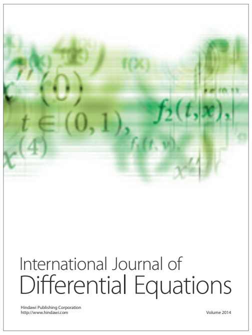

Advances in
Decision Sciences

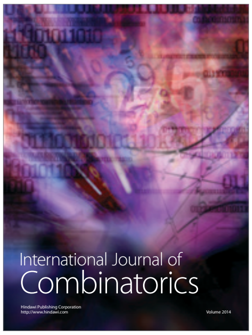

Mathematical Problems
in Engineering

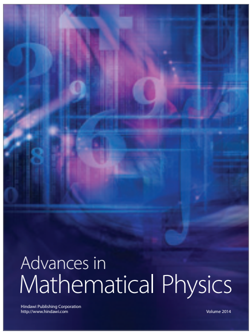

Algebra

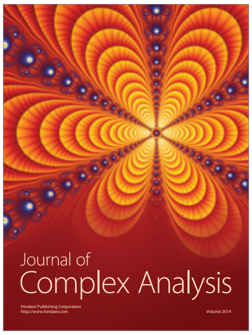

Journal of
Probability and Statistics

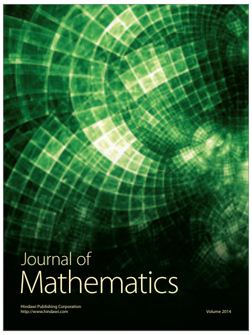
The Scientific
World Journal

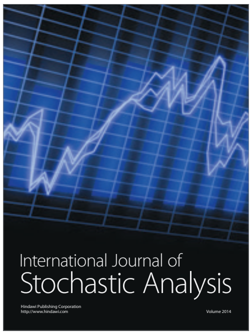

International Journal of
Differential Equations

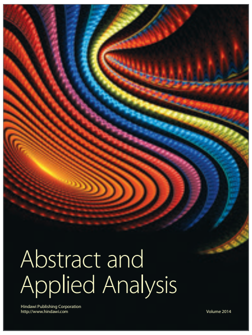

International Journal of
Combinatorics

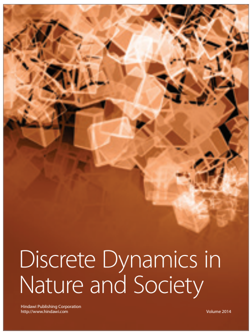

Advances in
Mathematical Physics

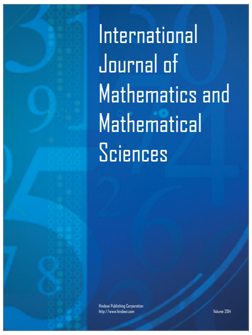

Journal of
Complex Analysis


Journal of
Mathematics


International Journal of
Stochastic Analysis

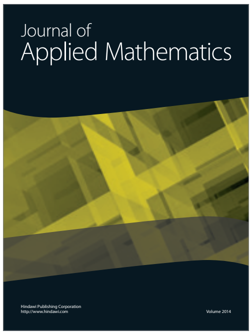

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


Journal of
Applied Mathematics

Journal of
Optimization