Convergence Analysis of Multiplicative Weight Noise Injection During Training

Kevin Ho
Department of Computer Science and Communication Engineering
Providence University
Sha-Lu, Taiwan
Email: ho@pu.edu.tw

Chi-sing Leung
Department of Electronic Engineering
City University of Hong Kong
Kowloon Tong, Hong Kong
Email: eeleungc@cityu.edu.hk

John Sum, Siu-chung Lau
Institute of Technology Management
National Chung Hsing University
Taichung, Taiwan.
Email:pfsum@nchu.edu.tw

***Abstract*—Injecting weight noise during training has been proposed for almost two decades as a simple technique to improve fault tolerance and generalization of a multilayer perceptron (MLP). However, little has been done regarding their convergence behaviors. Therefore, we presents in this paper the convergence proofs of two of these algorithms for MLPs. One is based on combining injecting multiplicative weight noise and weight decay (MWN-WD) during training. The other is based on combining injecting additive weight noise and weight decay (AWN-WD) during training. Let m be the number of hidden nodes of a MLP, α be the weight decay constant and S_b be the noise variance. It is showed that the convergence of MWN-WD algorithm is with probability one if $\alpha > \sqrt{S_b}m$. While the convergence of the AWN-WD algorithm is with probability one if $\alpha > 0$.**

***Keywords*-convergence; learning; MLP; weight noise;**

I. Introduction

To improve the fault tolerance of a multilayer perceptron (MLP), Murray & Edward [19], [20], [13] modified the conventional backpropagation training by injecting multiplicative weight noise during each step of training. By simulations on character encoder and eye-classifier problems, they found that the resultant multilayer perceptron has better tolerance ability against random weight fault and weight perturbation. Applying the same technique in real-time-recurrent-learning (RTRL), Jim *et al* [17] have also found that the generalization of a RNN can be improved. Moreover, the convergence speed is faster than conventional RTRL. While on-line weight noise injection training algorithms have succeeded in improving fault tolerance of a MLP, the generalization of a RNN, and convergence speed of training, not much analytical work has been done in regard to the (i) convergence proofs and (ii) objective functions of these algorithms.

Even the authors in [12], [20], [17] have only provided preliminary analyses on the effect of the prediction error of a neural network that is corrupted by weight noise (see Section II.C in [12], [20] for the analysis for MLP and see Section 3 in [17] for the analysis for RNN). G.An in [1] has attempted to these problems. In his paper, he considered three different on-line back-propagation training with noise injection. One of them is based on additive weight noise injection (see Section 4 in [1]). While his works in the other two algorithms are correct, his analysis on the case of weight noise injection is questionable. It is because he has not verified if the algorithm based on weight noise injection fulfils the conditions depicted in Bottou's Theorem [8]. By following the mathematics in [1], one can clearly figure out that the cost function derived by G. An is the prediction error of a MLP if it is corrupted by additive weight noise. It is not the corresponding objective function for on-line weight noise injection training algorithm for MLP.

Even though some other works have been done regarding the prediction error (or sensitivity analysis) of a MLPs [2], [3], [4], [5], [6], [12], [21], none of them worked on their convergence proofs. Until recently, we have showed that the convergence of injecting weight noise during training a RBF is with probability one [14], [16]. Nevertheless, we have showed that the objective function of injecting multiplicative weight noise (or additive weight noise) during training is essentially the mean square errors function. It means, injecting weight noise during training does not help to improve the fault tolerance or the generalization ability of a RBF. Unfortunately, our approach to the proof for RBF [16] cannot be applied to MLPs simply because Gladyshev Theorem is not applicable to MLPs.

After all, for almost fifteen years, the convergence proofs of these weight noise injection-based algorithms for MLP have yet been accomplished and their corresponding objective functions are still unknown.

Therefore, the primary focus of this paper is to analyze the convergence of these weight noise injection-based algorithms with application to MLPs. Two specific algorithms will be analyzed. The first one is based on combining multiplicative weight noise injection and weight decay during training. While the other is based on combining additive weight noise injection and weight decay during training. The main theorem we applied is the classical Doob's Martingale Convergence Theorem [11], [10].

The rest of the paper will present the main convergence theorems and the corresponding proofs for these weight noise injection-based algorithms for MLPs. In the next section, the background on the network model, the weight decay training algorithm will be described. Then in Section 3, the algorithms based on *combining weight noise injection and weight decay during training* will be summarized. Their corresponding objective functions will

be reviewed. In Section 4, the convergence of the algorithm based on combining multiplicative weight noise and weight decay during training will be proved. The convergence of the algorithm based on combining additive weight noise and weight decay during training will be proved in Section 5. Section 6 will prove that with probability one their weight vectors converge to local minimum of their corresponding objective functions. Conclusions are given in the last section.

II. Background

We assume that the training data set $\mathcal{D} = \{(\mathbf{x}_k, y_k)\}_{k=1}^N$ is generated by an unknown system, where $\mathbf{x}_k \in R^n$ is the input vector and $y_k \in R$ is the output.

A. Network Model

This unknown system is thus approximated by a MLP with n input nodes, m hidden nodes, and one linear output node, defined as follows :

$$f(\mathbf{x}_k, \mathbf{d}, \mathbf{A}, \mathbf{c}) = \mathbf{d}^T \mathbf{z}(\mathbf{A}^T \mathbf{x}_k + \mathbf{c}), \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_m] \in R^{n \times m}$ is the input-to-hidden weight matrix, $\mathbf{a}_i \in R^n$ is the input weight vector associated with the i^{th} hidden node, $\mathbf{c} = (c_1, \cdots, c_m)^T \in R^m$ is the input-to-hidden bias vector, $\mathbf{d} \in R^m$ is the hidden-to-output weight vector, and $\mathbf{z} = (z_1, \cdots, z_m)^T \in R^m$ is output vector of the hidden layer in which

$$z_i(\mathbf{x}_k, \mathbf{a}_i, c_i) = \frac{1}{1 + \exp(-(\mathbf{a}_i^T \mathbf{x}_k + c_i))} \tag{2}$$

for $i = 1, 2, \cdots, m$.

For the sake of presentation, we let $\mathbf{w}_i \in R^{(n+2)}$ be the parametric vector associated to the i^{th} hidden node, i.e.

$$\mathbf{w}_i = (d_i, \mathbf{a}_i^T, c_i)^T, \tag{3}$$

and $\mathbf{w} \in R^{m(n+2)}$ be a parametric vector augmenting all the parametric vectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m$. The output is denoted as $f(\mathbf{x}_k, \mathbf{w})$. Throughout the paper, we call $\mathbf{w}_1$, $\mathbf{w}_2, \cdots, \mathbf{w}_m$ and $\mathbf{w}$ the **weight vectors**.

Next, we let $\mathbf{g}(\mathbf{x}_k, \mathbf{w})$ be $\nabla_{\mathbf{w}} f(\mathbf{x}_k, \mathbf{w})$, where

$$\mathbf{g}(\mathbf{x}_k, \mathbf{w}) = (\nabla_{\mathbf{w}_1} f(\mathbf{x}_k, \mathbf{w})^T, \cdots, \nabla_{\mathbf{w}_m} f(\mathbf{x}_k, \mathbf{w})^T)^T.$$

As $\nabla_{\mathbf{w}_i} f(\mathbf{x}_k, \mathbf{w})$ depends entirely on $\mathbf{x}_k$ and $\mathbf{w}_i$, we denote it by $\mathbf{g}_i(\mathbf{x}_k, \mathbf{w}_i)$. Thus,

$$\mathbf{g}(\mathbf{x}_k, \mathbf{w}) = (\mathbf{g}_1(\mathbf{x}_k, \mathbf{w}_1)^T, \cdots, \mathbf{g}_m(\mathbf{x}_k, \mathbf{w}_m)^T)^T,$$

and

$$\mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i) = \begin{bmatrix} z_i \\ d_i z_i (1 - z_i) \mathbf{x}_k \\ d_i z_i (1 - z_i) \end{bmatrix}, \tag{4}$$

where $z_i = z_i(\mathbf{x}_k, \mathbf{a}_i, c_i)$.

If we let $\nabla_{\mathbf{w}} \mathbf{g}(\mathbf{x}, \mathbf{w}) \in R^{m(n+2) \times m(n+2)}$ be the Hessian matrix of $f(\mathbf{x}, \mathbf{w})$ with respect to the weight vector $\mathbf{w}$, one can readily show that

$$\nabla_{\mathbf{w}} \mathbf{g}(\mathbf{x}, \mathbf{w}) = \begin{bmatrix} \nabla_{\mathbf{w}_1} \mathbf{g}_1(\mathbf{x}, \mathbf{w}_1) & \cdots & \mathbf{0}_{(n+2)\times(n+2)} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{(n+2)\times(n+2)} & \cdots & \nabla_{\mathbf{w}_m} \mathbf{g}_m(\mathbf{x}, \mathbf{w}_m) \end{bmatrix}, \tag{5}$$

where

$$\nabla_{\mathbf{w}_i} \mathbf{g}_i(\mathbf{x}, \mathbf{w}) = \nabla \nabla_{\mathbf{w}_i} f(\mathbf{x}, \mathbf{w}) \tag{6}$$

for all $i = 1, \cdots, m$.

B. Weight Decay Training

In weight decay training, a sample is randomly drawn from the dataset $\mathcal{D}$ at each update step. We denote the sample being selected at the t^{th} step as $\{\mathbf{x}_t, y_t\}$. Once the input $\mathbf{x}_t$ has been fed in the MLP, the output is calculated by (1) and (2).

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{w}(t)) &= \mathbf{d}(t)^T \mathbf{z}(t) & (7) \\ \mathbf{z}(t) &= \mathbf{z}(\mathbf{A}(t)^T \mathbf{x}_t + \mathbf{c}(t)). & (8) \end{aligned}$$

By replacing $\mathbf{w}_i$ and $\mathbf{x}_k$ in (4) by $\mathbf{w}(t)$ and $\mathbf{x}_t$ respectively, the update equations for the weight vectors $\mathbf{w}_i$ (for $i = 1, 2, \cdots, m$) can thus be written as follows :

$$\begin{aligned} & \mathbf{w}_i(t+1) - \mathbf{w}_i(t) \\ = \; & \mu(t) \left\{ (y_t - f(\mathbf{x}_t, \mathbf{w}(t))) \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i(t)) - \alpha \mathbf{w}_i(t) \right\} \end{aligned} \tag{9}$$

where $\mu(t) > 0$ is the step size at the t^{th} step, and $\alpha > 0$ is the decay constant. The last term $-\alpha \mathbf{w}_i(t)$ in (9) sometimes is called forgetting term [18].

It has been proved in [**?**], [**?**] that the convergence of (9) is *with probability one* if $\sum_t \mu(t) = \infty$ and $\sum_t \mu(t)^2 < \infty$. However, we will show later in this paper that these conditions can be replaced by $\mu(t) \to 0$.

III. Combining weight noise injection and weight decay during training

Let $\mathbf{b}_1(t), \mathbf{b}_2(t), \cdots, \mathbf{b}_m(t) \in R^{(n+2)}$ be random vectors associated with the weight vectors $\mathbf{w}_1(t), \mathbf{w}_2(t), \cdots, \mathbf{w}_m(t)$ at step t. Elements in each random vector $\mathbf{b}_i(t)$ are independent mean zero Gaussian distributed random variables with variance denoted by S_b, i.e.

$$P(\mathbf{b}_i(t)) \sim \mathcal{N}\left(\mathbf{0}, S_b \mathbf{I}_{(n+2)\times(n+2)}\right) \tag{10}$$

for all $t \geq 0$. Furthermore, $\mathbf{b}_i(t_1)$ and $\mathbf{b}_i(t_2)$ are independent for $t_1 \neq t_2$.

We let $\tilde{\mathbf{w}}_i(t) = (\tilde{d}_i(t), \tilde{\mathbf{a}}_i(t)^T, \tilde{c}_i(t))^T$ be the perturbed weight vector associated with the i^{th} hidden node and the perturbed output of the i^{th} hidden node is denoted by $\tilde{z}_i(t) = z_i(\mathbf{x}_t, \tilde{\mathbf{a}}_i(t), \tilde{c}_i(t))$.

The update of $\mathbf{w}_i$ based on weight noise injection with weight decay can be written as follows :

$$\begin{aligned} & \mathbf{w}_i(t+1) - \mathbf{w}_i(t) \\ = \; & \mu(t)(y_t - f(\mathbf{x}_t, \tilde{\mathbf{w}}(t))) \mathbf{g}_i(\mathbf{x}_t, \tilde{\mathbf{w}}_i(t)) \\ & - \mu(t) \alpha \mathbf{w}_i(t). \end{aligned} \tag{11}$$

where $\tilde{\mathbf{w}}_i(t)$ is a perturbed weight vector and

$$\mathbf{g}_i(\mathbf{x}_t, \tilde{\mathbf{w}}_i(t)) = \begin{bmatrix} \tilde{z}_i(t) \\ \tilde{d}_i(t) \tilde{z}_i(t)(1 - \tilde{z}_i(t)) \mathbf{x}_t \\ \tilde{d}_i(t) \tilde{z}_i(t)(1 - \tilde{z}_i(t)) \end{bmatrix}, \tag{12}$$

$\mu(t) > 0$ is the step size at the t^{th} step, and $\alpha > 0$ is the decay constant.

A. MWN-WD algorithm

If the weight vector is perturbed by multiplicative weight noise, $\tilde{\mathbf{w}}_i(t)$ in (11) is given by

$$\tilde{\mathbf{w}}_i(t) = \mathbf{w}_i(t) + \mathbf{b}_i(t) \otimes \mathbf{w}_i(t), \tag{13}$$

where $\otimes$ is the elementwise multiplication operator defined as follows :

$$\mathbf{b}_i(t) \otimes \mathbf{w}_i(t) = (b_{i1}(t)w_{i1}(t), \cdots, b_{i(n+2)}(t)w_{i(n+2)}(t))^T. \tag{14}$$

As in [3], [4], [19], [20], the noise variance S_b is assumed to be a small positive value[1].

The output $f(\mathbf{x}_t, \tilde{\mathbf{w}}(t))$ and $\mathbf{g}_i(\mathbf{x}_t, \tilde{\mathbf{w}}_i(t))$ in (11) are approximated by

$$f(\mathbf{x}_t, \tilde{\mathbf{w}}) \approx f(\mathbf{x}_t, \mathbf{w}) + \sum_{i=1}^{m} \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i)^T (\mathbf{b}_i \otimes \mathbf{w}_i) \tag{15}$$

and

$$\mathbf{g}_i(\mathbf{x}_t, \tilde{\mathbf{w}}_i) \approx \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i) + \nabla_{\mathbf{w}_i} \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i) (\mathbf{b}_i \otimes \mathbf{w}_i), \tag{16}$$

where $\nabla_{\mathbf{w}_i} \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i)$ is given by (6). In (15) and (16) The parentheses (t) attached with $\mathbf{w}_i(t)$, $\tilde{\mathbf{w}}_i(t)$ and $\mathbf{b}_i(t)$ are omitted to save space.

Suppose each sample in the dataset $\mathcal{D}$ has equal probability to be selected. By (11), (15) and (16), the conditional expectation of $\mathbf{w}_i(t+1)$ over all random vectors $\mathbf{b}_1(t), \cdots, \mathbf{b}_m(t)$ on $\mathbf{w}(t)$ is given by :

$$E[\mathbf{w}_i(t+1)|\mathbf{w}(t)] = \mathbf{w}_i(t) + \mu(t)\mathbf{h}_i(\mathbf{w}(t)), \tag{17}$$

where

$$\begin{aligned} & \mathbf{h}_i(\mathbf{w}(t)) \\ = \; & \frac{1}{N}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}(t)))\mathbf{g}_i(\mathbf{x}_k, \mathbf{w}(t)) \\ & -\frac{S_b}{N}\sum_{k=1}^{N}\nabla_{\mathbf{w}_i}\mathbf{g}_i(\mathbf{x}_k, \mathbf{w}_i(t))\mathbf{v}_i(\mathbf{x}_k, \mathbf{w}_i(t)) \\ & -\alpha\mathbf{w}_i(t) \end{aligned} \tag{18}$$

and

$$\mathbf{v}_i(\mathbf{x}_k, \mathbf{w}_i(t)) = \mathbf{w}_i(t) \otimes \mathbf{w}_i(t) \otimes \mathbf{g}_i(\mathbf{x}_k, \mathbf{w}(t)). \tag{19}$$

We have showed that [16]

$$\mathbf{h}_i(\mathbf{w}(t)) = -\nabla_{\mathbf{w}_i} V(\mathbf{w}(t)), \tag{20}$$

[1]Note that this condition is not required in the convergence proof presented in Section 4.

where

$$\begin{aligned} V(\mathbf{w}) \; = \; & \frac{1}{2}\left\{\frac{1}{N}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}))^2 \right. \\ & \left. + \frac{S_b}{N}\sum_{k=1}^{N}\sum_{i=1}^{m}\left(\mathbf{w}_i^T\mathbf{g}_i(\mathbf{x}_k, \mathbf{w}_i)\right)^2\right\} \\ & -\frac{S_b}{N}\sum_{k=1}^{N}\int \mathbf{u}(\mathbf{x}_k, \mathbf{w})\mathbf{dw} \\ & +\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \end{aligned} \tag{21}$$

and

$$\mathbf{u}(\mathbf{x}_k, \mathbf{w}) = \begin{bmatrix} \mathbf{w}_1 \otimes \mathbf{g}_1(\mathbf{x}_k, \mathbf{w}_1) \otimes \mathbf{g}_1(\mathbf{x}_k, \mathbf{w}_1) \\ \mathbf{w}_2 \otimes \mathbf{g}_2(\mathbf{x}_k, \mathbf{w}_2) \otimes \mathbf{g}_2(\mathbf{x}_k, \mathbf{w}_2) \\ \vdots \\ \mathbf{w}_m \otimes \mathbf{g}_m(\mathbf{x}_k, \mathbf{w}_m) \otimes \mathbf{g}_m(\mathbf{x}_k, \mathbf{w}_m) \end{bmatrix} \tag{22}$$

The 3rd term in (21) is a line integral. In the later section, we will show that $\mathbf{w}(t)$ generated by (11) and (13) converges to a local minimum of this objective function $V(\mathbf{w})$.

B. AWN-WD algorithm

The definition of AWN-WD algorithm is similar to MWN-WD algorithm. The update equation is based on (11) but the perturbed weight vector is now given by

$$\tilde{\mathbf{w}}_i(t) = \mathbf{w}_i(t) + \mathbf{b}_i(t). \tag{23}$$

Thus, the output $f(\mathbf{x}_t, \tilde{\mathbf{w}}(t))$ and $\mathbf{g}_i(\mathbf{x}_t, \tilde{\mathbf{w}}_i)$ in (11) are approximated by

$$f(\mathbf{x}_t, \tilde{\mathbf{w}}) \approx f(\mathbf{x}_t, \mathbf{w}) + \sum_{i=1}^{m} \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i)^T\mathbf{b}_i, \tag{24}$$

and

$$\mathbf{g}_i(\mathbf{x}_t, \tilde{\mathbf{w}}_i) \approx \mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i) + \nabla_{\mathbf{w}_i}\mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i)\mathbf{b}_i, \tag{25}$$

where $\nabla_{\mathbf{w}_i}\mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i)$ is given by (6). Again, the parentheses (t) attached with $\mathbf{w}_i(t)$, $\tilde{\mathbf{w}}_i(t)$ and $\mathbf{b}_i(t)$ are omitted to save space.

Suppose each sample in the dataset $\mathcal{D}$ has equal probability to be selected. By (11), (24) and (25), the conditional expectation of $\mathbf{w}_i(t+1)$ over all random vectors $\mathbf{b}_1(t), \cdots, \mathbf{b}_m(t)$ on $\mathbf{w}(t)$ is given by :

$$E[\mathbf{w}_i(t+1)|\mathbf{w}(t)] = \mathbf{w}_i(t) + \mu(t)\mathbf{h}'_i(\mathbf{w}(t)), \tag{26}$$

where $\mathbf{h}'_i(\mathbf{w}(t))$

$$\begin{aligned} \mathbf{h}'_i(\mathbf{w}(t)) \; = \; & \frac{1}{N}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}(t)))g_i(\mathbf{x}_k, \mathbf{w}(t)) \\ & -\frac{S_b}{N}\sum_{k=1}^{N}\nabla_{\mathbf{w}_i}\mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i)\mathbf{g}_i(\mathbf{x}_k, \mathbf{w}(t)) \\ & -\alpha w_i(t). \end{aligned} \tag{27}$$

We have shown that [16]

$$\mathbf{h}_i'(\mathbf{w}(t)) = -\nabla_{\mathbf{w}_i} V'(\mathbf{w}(t)), \tag{28}$$

where

$$\begin{aligned} V'(\mathbf{w}) &= \frac{1}{2}\left\{\frac{1}{N}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k,\mathbf{w}))^2 \right. \\ &+\frac{S_b}{N}\sum_{k=1}^{N}\sum_{i=1}^{m}\|\mathbf{g}_i(\mathbf{x}_k,\mathbf{w})\|_2^2 \\ &\left.+\frac{\alpha}{2}\|\mathbf{w}\|_2^2\right\}. \end{aligned} \tag{29}$$

In the later section, we will show that $\mathbf{w}(t)$ generated by (11) and (23) converges to a local minimum of this objective function $V'(\mathbf{w})$.

IV. Convergence of MWN-WD algorithm

The convergence proof is conducted by the following steps. First, we consider the update of the output weight vector $\mathbf{d}(t)$ and apply Doob's Martingale Convergence Theorem to show that $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ exists. As the existence of $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ cannot imply the existence of $\lim_{t\to\infty}\mathbf{d}(t)$, we consider the update of the elements in $\mathbf{d}(t)$ and apply Doob's Martingale Convergence Theorem to show the existence of their limits.

The existence of $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ and $\lim_{t\to\infty} d_i(t)$ (for all $i = 1,\cdots,m$), together with the Doob's Martingale Convergence Theorem are then applied to show the existence of $\lim_{t\to\infty}\mathbf{a}_i(t)$ and $\lim_{t\to\infty} c_i(t)$. Finally, we conclude that $\lim_{t\to\infty}\mathbf{w}(t)$ exists.

Here and after, we let $\mathbf{b}_\mathbf{d}(t)$ be the random vector associated with the output vector $\mathbf{d}$. That is,

$$\mathbf{b}_\mathbf{d}(t) = (b_{11}(t), b_{21}(t),\cdots,b_{m1}(t))^T, \tag{30}$$

where $b_{i1}(t)$ is the first element in $\mathbf{b}_i(t)$. Besides, we use the notation $E_\mathbf{d}[\cdot|\mathbf{w}(t)]$ denoting the conditional expectation that is taken over the random vector $\mathbf{b}_\mathbf{d}(t)$ only.

A. Existence of $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$

By (1), (11) and (12), the update of $d_i(t)$ can be expressed as follows :

$$\begin{aligned} & d_i(t+1) - d_i(t) \\ =\ & \mu(t)\left\{(y_t - \tilde{\mathbf{d}}^T(t)\tilde{\mathbf{z}}(t))\tilde{z}_i(t) - \alpha\, d_i(t)\right\}. \end{aligned} \tag{31}$$

In vector-matrix form,

$$\begin{aligned} & \mathbf{d}(t+1) - \mathbf{d}(t) \\ =\ & \mu(t)\left\{(y_t - \tilde{\mathbf{d}}^T(t)\tilde{\mathbf{z}}(t))\tilde{\mathbf{z}}(t) - \alpha\,\mathbf{d}(t)\right\}. \end{aligned} \tag{32}$$

Based on (32), the boundedness of $E[\|\mathbf{d}(t)\|_2]$ and the existence of $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ can be stated in the following lemma.

Lemma 1: For the algorithm based on (11) and (13), if $0 < \mu(t)(\alpha - \sqrt{S_b}m) < 1$ for all $t \geq 0$, then $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ exists and is finite with probability one,

Proof: We rewrite the update of $\mathbf{d}(t)$ given by (32) as follows :

$$\begin{aligned} \mathbf{d}(t+1) &= (1-\mu(t)\alpha)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t) \\ &-\mu(t)\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\mathbf{d}(t) \\ &-\mu(t)\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\left(\mathbf{b}_\mathbf{d}(t)\otimes\mathbf{d}(t)\right). \end{aligned} \tag{33}$$

Here, we let

$$\mathbf{B}(t) = (1-\mu(t)\alpha)I_{m\times m} - \mu(t)\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t). \tag{34}$$

Equation (33) can be written as follows :

$$\begin{aligned} \mathbf{d}(t+1) &= \mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t) \\ &-\mu(t)\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\left(\mathbf{b}_\mathbf{d}(t)\otimes\mathbf{d}(t)\right). \end{aligned} \tag{35}$$

Since the elements in $\mathbf{b}_\mathbf{d}(t)$ are identical and independent mean zero Gaussian random variables with variance S_b,

$$\begin{aligned} & E_\mathbf{d}\left[\left(\mathbf{b}_\mathbf{d}(t)\otimes\mathbf{d}(t)\right)\left(\mathbf{b}_\mathbf{d}(t)\otimes\mathbf{d}(t)\right)^T|\mathbf{w}(t)\right] \\ =\ & S_b\begin{bmatrix} d_1(t)^2 & 0 & \cdots & 0 \\ 0 & d_2(t)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_m(t)^2 \end{bmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} & E_\mathbf{d}[\|\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\left(\mathbf{b}_\mathbf{d}(t)\otimes\mathbf{d}(t)\right)\|_2^2|\mathbf{w}(t)] \\ =\ & S_b\|\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\mathbf{d}(t)\|_2^2, \end{aligned} \tag{36}$$

where **Tr** is the trace operator.

Given $\mathbf{w}(t)$, the expectation of the $\|\mathbf{d}(t+1)\|_2^2$ over the random vector $\mathbf{b}_\mathbf{d}(t)$ is given by

$$\begin{aligned} & E_\mathbf{d}[\|\mathbf{d}(t+1)\|_2^2|\mathbf{w}(t)] \\ =\ & \|\mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t)\|_2^2 + \mu(t)^2S_b\|\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\mathbf{d}(t)\|_2^2 \\ \leq\ & \|\mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t)\|_2^2 + \mu(t)^2S_bm^2\|\mathbf{d}(t)\|_2^2 \\ \leq\ & \left(\|\mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t)\|_2 + \mu(t)\sqrt{S_b}m\|\mathbf{d}(t)\|_2\right)^2 \end{aligned} \tag{37}$$

The last inequality based on the fact that the eigenvalues of $\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)$ are 0 and $\sum_{i=1}^m\tilde{z}_i(t)^2$.

Further by Jensen Inequality that $E_\mathbf{d}[\|\mathbf{d}(t+1)\|_2|\mathbf{w}(t)] \leq \left(E_\mathbf{d}[\|\mathbf{d}(t+1)\|_2^2|\mathbf{w}(t)]\right)^{1/2}$, and then by Triangle Inequality,

$$\begin{aligned} & E_\mathbf{d}[\|\mathbf{d}(t+1)\|_2|\mathbf{w}(t)] \\ \leq\ & \|\mathbf{B}(t)\mathbf{d}(t)\|_2 + \mu(t)\|y_t\tilde{\mathbf{z}}(t)\|_2 + \mu(t)\sqrt{S_b}m\|\mathbf{d}(t)\|_2 \\ \leq\ & (1-\mu(t)(\alpha-\sqrt{S_b}m))\|\mathbf{d}(t)\|_2 + \mu(t)\|y_t\tilde{\mathbf{z}}(t)\|_2 \end{aligned} \tag{38}$$

The last inequality based on the fact that the eigenvalues $\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)$ are 0 and $\sum_{i=1}^m\tilde{z}_i(t)^2$. Hence, the eigenvalue of $\mathbf{B}(t)$ must be less than or equal to $(1-\mu(t)\alpha)$.

To save space, we let $\alpha' = \alpha - \sqrt{S_b}m$. As y_t is generally bounded for all $t \geq 0$, $\|y_t\tilde{\mathbf{z}}(t)\|_2$ is bounded by a positive constant. Let it be κ_d. Thus,

$$E_\mathbf{d}[\|\mathbf{d}(t+1)\|_2|\mathbf{w}(t)] \leq (1-\mu(t)\alpha')\|\mathbf{d}(t)\|_2 + \mu(t)\kappa_d. \tag{39}$$

As the right hand side of (39) is independent of the random vector $\mathbf{b}(t)$,

$$E[\|\mathbf{d}(t+1)\|_2|\mathbf{w}(t)] \leq (1-\mu(t)\alpha')\|\mathbf{d}(t)\|_2+\mu(t)\kappa_d. \quad (40)$$

Equivalently,

$$\begin{aligned} & E[\|\mathbf{d}(t+1)\|_2 - \kappa_d/\alpha'|\mathbf{w}(t)] \\ \leq\ & (1-\mu(t)\alpha')\left(\|\mathbf{d}(t)\|_2 - \kappa_d/\alpha'\right). \end{aligned} \quad (41)$$

Let

$$\beta(t) = \|\mathbf{d}(t)\|_2 - \frac{\kappa_d}{\alpha'}. \quad (42)$$

It is clear that $\{\beta(t)\}_{t\geq 0}$ is a supermartingale and

$$E[|\beta(t)|] \leq E[|\beta(t-1)|] \leq \cdots \leq E[|\beta(0)|]. \quad (43)$$

By Doob's Martingale Convergence Theorem, $\lim_{t\to\infty}\beta(t)$ exists and is finite with probability one. Then from (42), $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ exists and is fine with probability one. The proof is completed. **Q.E.D**

Lemma 1 is crucial for the following proofs on the existence of $\lim_{t\to\infty}\mathbf{d}(t)$, $\lim_{t\to\infty}\mathbf{a}_i(t)$ and $\lim_{t\to\infty}c_i(t)$. The idea can be described in the rest of this subsection.

As $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ exits with probability one, $\lim_{t\to\infty}E[\|\mathbf{d}(t+1)\|_2|\mathbf{w}(t)] = \lim_{t\to\infty}\|\mathbf{d}(t+1)\|_2$. Thus, for any small positive ϵ, there exists a time t^* such that

$$P(|\|\mathbf{d}(t)\|_2 - \kappa_d/\alpha'| > \epsilon) = 0 \quad (44)$$

for all $t \geq t^*$ or equivalently

$$\|\mathbf{d}(t)\|_2 < \kappa_d/\alpha' + \epsilon \quad (45)$$

is with probability one for all $t \geq t^*$.

Making use of (44), we can therefore derive inequalities bounding $E[d_i(t+1)|\mathbf{w}(t)]$, $E[a_{ij}(t+1)|\mathbf{w}(t)]$ and $E[c_i(t+1)|\mathbf{w}(t)]$ for $t \geq t^*$, for $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n$.

Then, we can define supermartingales for each $d_i(t)$, $a_{ij}(t)$ and $c_i(t)$ respectively and show the existence of their limits by Doob's Martingale Convergence Theorem.

B. Existence of $\lim_{t\to\infty}\mathbf{d}(t)$

To show the existence of the limit of $\mathbf{d}(t)$, we consider (31) and $t \geq t^*$.

$$\begin{aligned} E_{\mathbf{d}}[d_i(t+1)|\mathbf{w}(t)] &= (1-\mu(t)\alpha)d_i(t) \\ &\quad +\mu(t)(y_t - \mathbf{d}^T(t)\tilde{\mathbf{z}}(t))\tilde{z}_i(t) \\ &\leq (1-\mu(t)\alpha)d_i(t) + \mu(t)\kappa_d \\ &\quad +\mu(t)\|\mathbf{d}(t)\|_2. \end{aligned} \quad (46)$$

As a result of (44),

$$\begin{aligned} E_{\mathbf{d}}[d_i(t+1)|\mathbf{w}(t)] &\leq (1-\mu(t)\alpha)d_i(t) + \mu(t)\kappa_d \\ &\quad +\mu(t)(\kappa_d/\alpha' + \epsilon). \end{aligned} \quad (47)$$

Since the right hand side is independent of other random variables in $\mathbf{b}(t)$, we can write that

$$\begin{aligned} E[d_i(t+1)|\mathbf{w}(t)] &\leq (1-\mu(t)\alpha)d_i(t) + \mu(t)\kappa_d \\ &\quad +\mu(t)(\kappa_d/\alpha' + \epsilon) \end{aligned} \quad (48)$$

for all $t \geq t^*$.

Hence, we can define a random process $\{\gamma(t)\}_{t\geq 0}$ in which

$$\gamma(t) = d_i(t+t^*) - \frac{\kappa_d + \kappa_d/\alpha' + \epsilon}{\alpha} \quad (49)$$

and clearly $E[|\gamma(t)||\mathbf{w}(t^*)] \leq \cdots \leq E[|\gamma(0)||\mathbf{w}(t^*)]$. Therefore, by Doob's Martingale Convergence Theorem, $\lim_{t\to\infty}\gamma(t)$ exists and is finite with probability one. We can conclude that $\lim_{t\to\infty}d_i(t)$ exists and is finite with probability one. As the same procedure applies to all $i = 1, 2, \cdots, m$, we can have the following lemma.

Lemma 2: For the algorithm based on (11) and (13), if $0 < \mu(t)(\alpha - \sqrt{S_b}m) < 1$ *for all* $t \geq 0$*, then* $\lim_{t\to\infty}\mathbf{d}(t)$ *exists and its elements are finite with probability one,*

C. Existence of $\lim_{t\to\infty}\mathbf{a}_i(t)$

The proof of the existence of $\lim_{t\to\infty}\mathbf{a}_i(t)$ is similar that of the proof of Lemma 2. By (1), (11) and (12), the update of $\mathbf{a}_i(t)$ can be expressed as follows :

$$\begin{aligned} & \mathbf{a}_i(t+1) \\ =\ & (1-\mu(t)\alpha)\mathbf{a}_i(t) + \mu(t)y_t\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)\mathbf{x}_t \\ & -\mu(t)\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)\mathbf{x}_t\tilde{\mathbf{z}}^T(t)\tilde{\mathbf{d}}(t). \end{aligned} \quad (50)$$

Note from (13) and (30) that

$$\tilde{d}_i(t) = d_i(t) + b_{i1}(t)d_i(t) \quad (51)$$

and

$$\tilde{d}_i(t)\tilde{\mathbf{d}}(t) = (d_i(t) + b_{i1}(t)d_i(t))\left(\mathbf{b}_{\mathbf{d}}(t)\otimes\mathbf{d}(t)\right). \quad (52)$$

Let us consider the j^{th} element in $\mathbf{a}_i(t)$.

$$\begin{aligned} & a_{ij}(t+1) \\ =\ & (1-\mu(t)\alpha)a_{ij}(t) + \mu(t)y_t\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)x_{tj} \\ & -\mu(t)\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)x_{tj}\tilde{\mathbf{z}}^T(t)\tilde{\mathbf{d}}(t). \end{aligned} \quad (53)$$

Lemma 3: For the algorithm based on (11) and (13), if $0 < \mu(t)(\alpha - \sqrt{S_b}m) < 1$ *for all* $t \geq 0$*, then for all* $i = 1, 2\cdots, m$*,* $\lim_{t\to\infty}\mathbf{a}_i(t)$ *exists and its elements are finite with probability one.*

Proof: Given $\mathbf{w}(t)$ and taking expectation of (53) over $\mathbf{b}_{\mathbf{d}}(t)$, $E_{\mathbf{d}}[a_{ij}(t+1)|\mathbf{w}(t)]$ can be expressed as follows :

$$\begin{aligned} & E_{\mathbf{d}}[a_{ij}(t+1)|\mathbf{w}(t)] \\ =\ & (1-\mu(t)\alpha)a_{ij}(t) + \mu(t)y_tv_1(t)d_i(t) \\ & -\mu(t)v_1(t)d_i(t)\tilde{\mathbf{z}}^T(t)(\mathbf{d}(t) + S_bd_i(t)\mathbf{e}_i), \end{aligned} \quad (54)$$

where

$$v_1(t) = \tilde{z}_i(t)(1-\tilde{z}_i(t))x_{tj} \quad (55)$$

and

$$\mathbf{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T. \quad (56)$$

Again, for $t \geq t^*$, by (44) and (54),

$$\begin{aligned} & E_{\mathbf{d}}[a_{ij}(t+1)|\mathbf{w}(t)] \\ \leq\ & (1-\mu(t)\alpha)a_{ij}(t) + \mu(t)|d_i(t)||x_{tj}| \\ & +\mu(t)|x_{tj}|\left(|d_i(t)|\|\mathbf{d}(t)\|_2 + S_bd_i(t)^2\right). \end{aligned} \quad (57)$$

Since $|x_{tj}|$ is bounded, say by κ_x, we can replace the second term and the third term by $\mu(t)\kappa_a$, where

$$\kappa_a = (\kappa_d/\alpha' + \epsilon)^2 \kappa_x (1 + S_b). \quad (58)$$

Thus,

$$E_{\mathbf{d}}[a_{ij}(t+1)|\mathbf{w}(t)] \leq (1-\mu(t)\alpha)a_{ij}(t) + \mu(t)\kappa_a. \quad (59)$$

As the right hand side is independent of other random variables,

$$E[a_{ij}(t+1)|\mathbf{w}(t)] \leq (1-\mu(t)\alpha)a_{ij}(t) + \mu(t)\kappa_a. \quad (60)$$

Similar to Lemma 2, we can define a random process $\{\xi(t)\}_{t\geq 0}$ as follows : $\xi(t) = a_{ij}(t+t^*) - \frac{\kappa_a}{\alpha}$ for all $t \geq t^*$ and clearly $E[|\xi(t)||\mathbf{w}(t^*)] \leq \cdots \leq E[|\xi(0)||\mathbf{w}(t^*)]$. Therefore, by Doob's Martingale Convergence Theorem, $\lim_{t\to\infty}\xi(t)$ exists and is finite with probability one. We can conclude that $\lim_{t\to\infty} a_{ij}(t)$ exists and is finite with probability one. Thus, for all $i = 1, \cdots, m$, $\lim_{t\to\infty}\mathbf{a}_i(t)$ exists and its elements are finite with probability one.

D. Existence of $\lim_{t\to\infty}|c_i(t)|$

By (1), (11) and (12), the update of $c_i(t)$ can be expressed as follows :

$$\begin{aligned} & c_i(t+1) \\ = \; & (1-\mu(t)\alpha)c_i(t) + \mu(t)y_t\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t) \\ & -\mu(t)\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)\tilde{\mathbf{z}}^T(t)\tilde{\mathbf{d}}(t). \end{aligned} \quad (61)$$

Suppose, we define two augmented vectors as that

$$\mathbf{x}'_t = \begin{bmatrix} \mathbf{x}_t \\ 1 \end{bmatrix} \text{ and } \mathbf{a}'_i(t) = \begin{bmatrix} \mathbf{a}_i(t) \\ c_i(t) \end{bmatrix}.$$

We can combine (50) and (61) together and come up with the following update equation.

$$\begin{aligned} & \mathbf{a}'_i(t+1) \\ = \; & (1-\mu(t)\alpha)\mathbf{a}_i(t) + \mu(t)y_t\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)\mathbf{x}'_t \\ & -\mu(t)\tilde{z}_i(t)(1-\tilde{z}_i(t))\tilde{d}_i(t)\mathbf{x}^*_t\tilde{\mathbf{z}}^T(t)\tilde{\mathbf{d}}(t). \end{aligned} \quad (62)$$

Repeating the same steps as the proof of Lemma 3, we can conclude the existence of $\lim_{t\to\infty}\mathbf{a}'_i(t)$ and thus $\lim_{t\to\infty} c_i(t)$ is with probability one.

Lemma 4: For the algorithm based on (11) and (13), if $0 < \mu(t)(\alpha - \sqrt{S_b}m) < 1$ for all $t \geq 0$, then for all $i = 1, 2\cdots, m$, $\lim_{t\to\infty} c_i(t)$ exists and is finite with probability one.

E. Existence of $\lim_{t\to\infty}\mathbf{w}(t)$

As a direct implication from Lemma 2-4, we state without proof the following theorem for the weight vector $\mathbf{w}(t)$.

Theorem 1: For the algorithm based on (11) and (13), if $0 < \mu(t)(\alpha - \sqrt{S_b}m) < 1$ for all $t \geq 0$, then $\lim_{t\to\infty}\mathbf{w}(t)$ exists and its elements are finite with probability one.

Let us define a bounded region $\Omega_{\bar{\epsilon}}(\mathbf{w}^*)$ which is centered at $\mathbf{w}^*$ and $\|\mathbf{w} - \mathbf{w}^*\| \leq \bar{\epsilon}$ for all $\mathbf{w} \in \Omega_{\bar{\epsilon}}(\mathbf{w}^*)$. Theorem 1 implies that for any arbitrary small positive $\bar{\epsilon}$, there must exist a bounded region $\Omega_{\bar{\epsilon}}(\mathbf{w}^*)$ and a time $\bar{t}(\mathbf{w}^*)$, such that for all $t \geq \bar{t}(\mathbf{w}^*)$

$$P(\mathbf{w}(t) \in \Omega_{\bar{\epsilon}}(\mathbf{w}^*)) = 1. \quad (63)$$

This final equation is very useful in the subsequent analysis.

V. Convergence of AWN-WD algorithm

Basically, the steps of proof for the AWN-WD algorithm are the same as the proof for the MWN-WD algorithm. The only difference is in the definition of $\tilde{\mathbf{w}}$. Owing to save space, we skip some of the proofs in this section. Only the existence of $\lim_{t\to\infty}\|\mathbf{d}(t)\|_2$ is proved, as it is the key step to show that noise variance S_b does not affect the convergence.

Theorem 2: For the algorithm based on (11) and (23), if $0 < \mu(t)\alpha < 1$ for all $t \geq 0$, then $\lim_{t\to\infty}\mathbf{w}(t)$ exists and its elements are finite with probability one.

Proof: Replace $\mathbf{b}_{\mathbf{d}}(t) \otimes \mathbf{d}(t)$ in (33) and (35) by $\mathbf{b}_{\mathbf{d}}(t)$, the update of $\mathbf{d}(t)$ is given by

$$\mathbf{d}(t+1) = \mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t) - \mu(t)\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\mathbf{b}_{\mathbf{d}}(t). \quad (64)$$

Note that $\tilde{\mathbf{z}}(t)$ in (64) is now depended on $\mathbf{w}(t) + \mathbf{b}(t)$ instead of $\mathbf{w}(t) + \mathbf{b}(t) \otimes \mathbf{w}(t)$.

Given $\mathbf{w}(t)$, the expectation of the $\|\mathbf{d}(t+1)\|_2^2$ over the random vector $\mathbf{b}_{\mathbf{d}}(t)$ is then given by

$$\begin{aligned} & E_{\mathbf{d}}[\|\mathbf{d}(t+1)\|_2^2|\mathbf{w}(t)] \\ = \; & \|\mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t)\|_2^2 \\ & +\mu(t)^2 E_{\mathbf{d}}[\mathbf{b}_{\mathbf{d}}^T(t)\left(\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\right)^2\mathbf{b}_{\mathbf{d}}(t)|\mathbf{w}(t)]. \end{aligned} \quad (65)$$

As,

$$\begin{aligned} & E_{\mathbf{d}}[\mathbf{b}_{\mathbf{d}}^T(t)\left(\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\right)^2\mathbf{b}_{\mathbf{d}}(t)|\mathbf{w}(t)] \\ = \; & \mathbf{Tr}\left\{E_{\mathbf{d}}[\left(\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\right)^2\mathbf{b}_{\mathbf{d}}(t)\mathbf{b}_{\mathbf{d}}^T(t)|\mathbf{w}(t)]\right\} \\ = \; & S_b\mathbf{Tr}\left\{\tilde{\mathbf{z}}(t)\underbrace{\tilde{\mathbf{z}}^T(t)\tilde{\mathbf{z}}(t)}_{\|\tilde{\mathbf{z}}(t)\|_2^2}\tilde{\mathbf{z}}^T(t)\right\} \\ = \; & S_b\|\tilde{\mathbf{z}}(t)\|_2^2\mathbf{Tr}\left\{\tilde{\mathbf{z}}(t)\tilde{\mathbf{z}}^T(t)\right\} \\ = \; & S_b\|\tilde{\mathbf{z}}(t)\|_2^4 \\ \leq \; & S_b m^2. \end{aligned} \quad (66)$$

The last inequality is due to the fact that the elements in $\tilde{\mathbf{z}}(t)$ are all in between 0 and 1. $\|\tilde{\mathbf{z}}(t)\|_2 \leq \sqrt{m}$. As a result,

$$\begin{aligned} & E_{\mathbf{d}}[\|\mathbf{d}(t+1)\|_2^2|\mathbf{w}(t)] \\ = \; & \|\mathbf{B}(t)\mathbf{d}(t) + \mu(t)y_t\tilde{\mathbf{z}}(t)\|_2^2 + \mu(t)^2 S_b m^2. \end{aligned} \quad (67)$$

By Jensen Inequality and Triangle Inequality,

$$\begin{aligned} & E_{\mathbf{d}}[\|\mathbf{d}(t+1)\|_2|\mathbf{w}(t)] \\ \leq \; & \|\mathbf{B}(t)\mathbf{d}(t)\|_2 + \mu(t)\|y_t\tilde{\mathbf{z}}(t)\|_2 + \mu(t)\sqrt{S_b}m \\ \leq \; & (1-\mu(t)\alpha)\|\mathbf{d}(t)\|_2 + \mu(t)\|y_t\tilde{\mathbf{z}}(t)\|_2 \\ & +\mu(t)\sqrt{S_b}m. \end{aligned} \quad (68)$$

As y_t is bounded for all $t \geq 0$, $\|y_t \tilde{\mathbf{z}}(t)\|_2$ is bounded by a positive constant. We can then let κ'_d be the bound of $\|y_t \tilde{\mathbf{z}}(t)\|_2 + \sqrt{S_b} m$.

We can then replace α' by α and κ_d by κ'_d in the steps from (39) to (43) and conclude that if $0 < \mu(t)\alpha < 1$, $\lim_{t\to\infty} \|\mathbf{d}(t)\|_2$ exists and is fine with probability one.

As the proofs for the existence of $\lim_{t\to\infty} \mathbf{d}(t)$, $\lim_{t\to\infty} \mathbf{a}_i(t)$ and $\lim_{t\to\infty} c_i(t)$ following similar steps as the proof for MWN-WD algorithm, their proofs are skipped. **Q.E.D.**

One should note that the condition for the convergence of AWN-WD algorithm does not depend the value of S_b. It depends on the values of the step size $\mu(t)$ and the decay constant α only.

VI. Asymptotic Properties of $V(\mathbf{w}^*)$

While Theorem 1 and Theorem 2 state the existence of $\mathbf{w}(t)$ when $t \to \infty$, they could not imply that their limits are located in local minimum of the corresponding objective functions. Therefore, it is necessarily to show that their locations are at local minimum. As the steps of proofs for both the MWN-WD algorithm and AWN-WD algorithm are the same, only the theorem and the proof regarding the MWN-WD algorithm will be presented in this section. The theorem regarding the AWN-WD algorithm will be stated without proof.

Before proceed to the statement of theorem, we need to make three more assumptions on the noise variance S_b and the step size $\mu(t)$ as follows :

$$S_b \ll 1, \tag{69}$$

$$\mu(t) \to 0 \quad \text{for all} \quad t \geq 0, \tag{70}$$

$$\sum_{\tau=t}^{\infty} \mu(\tau) = \infty \quad \text{for all} \quad t \geq 0. \tag{71}$$

The first assumption on S_b a common assumption made by other researchers [3], [4], [19], [20]. With this assumption, the approximations for $f(\mathbf{x}_t, \mathbf{w}(t))$ (15) and $\mathbf{g}_i(\mathbf{x}_t, \mathbf{w}_i(t))$ (16) will be making sense. The objective function (21) is in simple close form.

The second assumption is owing to simplify the approximation of $V(\mathbf{w}(t+1)) - V(\mathbf{w}(t))$ by ignoring higher order terms containing $\mu(t)^2$. The third assumption is to ensure that $\sum_{\tau=t}^{\infty} \mu(\tau) \|\nabla_{\mathbf{w}} V(\mathbf{w}(\tau))\|_2^2$ diverges if $\|\nabla_{\mathbf{w}} V(\mathbf{w}(t))\|_2$ does not converge.

With the assumptions on (69), (70) and (71), the property of $\nabla_{\mathbf{w}} V(\mathbf{w}^*)$ can then be stated as the following theorem.

Theorem 3: For the algorithm based on (11) and (13), if (i) $(\alpha - \sqrt{S_b} m) > 0$, (ii) $S_b \ll 1$, (iii) $\mu(t) \to 0$ for all $t \geq 0$ and (iv) $\sum_{\tau=t}^{\infty} \mu(t) = \infty$ for any $t \geq 0$, then $\mathbf{w}(t)$ converges to the location in which

$$\nabla_{\mathbf{w}} V(\mathbf{w}^*) = \lim_{t\to\infty} \nabla_{\mathbf{w}} V(\mathbf{w}(t)) = \mathbf{0}, \tag{72}$$

where $V(\mathbf{w})$ is a scalar function given by (21).

Proof: First of all, as $\mathbf{h}_i(\mathbf{w})$ is the gradient vector of $V(\mathbf{w})$ with differentiable functional elements,

$$V(\mathbf{w}) \text{ is differentiable for all } \mathbf{w} \text{ and} \tag{73}$$

$$\nabla_{\mathbf{w}} V(\mathbf{w}) \text{ is differentiable for all } \mathbf{w}. \tag{74}$$

From Condition (i) and (iii), $\lim_{t\to\infty} \mathbf{w}(t)$ exists and with finite elements. By virtue of Theorem 1 and (73), $\lim_{t\to\infty} V(\mathbf{w}(t))$ and $\lim_{t\to\infty} \nabla_{\mathbf{w}} V(\mathbf{w}(t))$ exist with probability one.

It implies that for any arbitrary ϵ_V, there exists a time t_V, such that

$$P(|V(\mathbf{w}^*) - V(\mathbf{w}(t))| \leq \epsilon_V) = 1 \tag{75}$$

for all $t \geq t_V$.

By Taylor expansion of $V(\mathbf{w}(t+1))$ about $\mathbf{w}(t)$,

$$V(\mathbf{w}(t+1)) = V(\mathbf{w}(t)) + (\nabla_{\mathbf{w}} V(\mathbf{w}(t)))^T \Delta\mathbf{w}(t), \tag{76}$$

where

$$\Delta\mathbf{w}(t) = \mu(t)(y_t - \tilde{\mathbf{d}}^T(t)\tilde{\mathbf{z}}(t))\mathbf{g}(\mathbf{x}_t, \mathbf{w}(t)), \tag{77}$$

in which $\mathbf{g} = (\mathbf{g}_1^T, \mathbf{g}_2^T, \cdots, \mathbf{g}_m^T)^T$. By (17) and (18),

$$E[V(\mathbf{w}(t+1))|\mathbf{w}(t)] = V(\mathbf{w}(t)) - \mu(t) \|\nabla_{\mathbf{w}} V(\mathbf{w}(t))\|^2. \tag{78}$$

Hence,

$$\begin{aligned} & E[V(\mathbf{w}^*) - V(\mathbf{w}(t_V))|\mathbf{w}(t_V)] \\ = & -\sum_{\tau=t_V}^{\infty} \mu(\tau) E\left[\|\nabla_{\mathbf{w}} V(\mathbf{w}(\tau))\|^2 \Big| \mathbf{w}(t_V) \right]. \end{aligned} \tag{79}$$

From (75), the right hand side of (79) must be with probability one smaller than ϵ_V.

$$\sum_{t=t_V}^{\infty} \mu(t) E\left[\|\nabla_{\mathbf{w}} V(\mathbf{w}(t))\|^2 \Big| \mathbf{w}(t_V) \right] \leq \epsilon_V. \tag{80}$$

By virtue of (71) and the inequality that $E[\|q\|] \leq (E[\|q\|_2^2])^{1/2}$, it can be proved by contradiction that

$$E\left[\|\nabla_{\mathbf{w}} V(\mathbf{w}^*)\| \,|\, \mathbf{w}(t_V) \right] = 0. \tag{81}$$

As $\|\nabla_{\mathbf{w}} V(\mathbf{w}^*)\|$ is non-negative, $\|\nabla_{\mathbf{w}} V(\mathbf{w}^*)\| = 0$. Hence

$$\nabla_{\mathbf{w}} V(\mathbf{w}^*) = \mathbf{0}. \tag{82}$$

The proof is completed. **Q.E.D.**

For the AWN-WD algorithm, we let $\mathbf{w}^{**}$ be the limit $\lim_{t\to\infty} \mathbf{w}(t)$. The the property of $\nabla_{\mathbf{w}} V'(\mathbf{w}^{**})$ is stated without proof as the following theorem.

Theorem 4: For the algorithm based on (11) and (23), if (i) $\alpha > 0$, (ii) $S_b \ll 1$, (iii) $\mu(t) \to 0$ for all $t \geq 0$ and (iv) $\sum_{\tau=t}^{\infty} \mu(t) = \infty$ for any $t \geq 0$, then $\mathbf{w}(t)$ converges to the location in which

$$\nabla_{\mathbf{w}} V'(\mathbf{w}^{**}) = \lim_{t\to\infty} \nabla_{\mathbf{w}} V'(\mathbf{w}(t)) = \mathbf{0}, \tag{83}$$

where $V'(\mathbf{w})$ is a scalar function given by (29).

VII. Conclusion

In this paper, we have presented two training algorithms based on on-line combining weight noise injection and weight decay. Their algorithms and objective functions have been presented. Their convergence are proved. Apart from the convergence proof, we have also showed that the locations the weight vectors converge to the local minimum of the corresponding objective functions.

Acknowledgement

The research work reported in this paper is supported in part by Taiwan National Science Council (NSC) Research Grant 97-2221-E-005-050 and 98-2221-E-005-048.

References

[1] An G. The effects of adding noise during backpropagation training on a generalization performance, *Neural Computation*, Vol.8, 643-674, 1996.

[2] Basalyga G. and E. Salinas, When response variability increases neural network robustness to synaptic noise, *Neural Computation*, Vol.18, 1349-1379, 2006.

[3] Bernier J.L. *et al*, Obtaining fault tolerance multilayer perceptrons using an explicit regularization, *Neural Processing Letters*, Vol.12, 107-113, 2000.

[4] Bernier J.L. *et al*, A quantitative study of fault tolerance, noise immunity and generalization ability of MLPs, *Neural Computation*, Vol.12, 2941-2964, 2000.

[5] Bernier J.L. *et al*, Improving the tolerance of multilayer perceptrons by minimizing the statistical sensitivity to weight deviations, *Neurocomputing*, Vol.31, 87-103, 2000.

[6] Bernier J.L. *et al*, Assessing the noise immunity and generalization of radial basis function networks, *Neural Processing Letter*, Vol.18(1), 35-48, 2003.

[7] Bishop C.M., Training with noise is equivalent to Tikhonov regularization, *Neural Computation*, Vol.7, 108-116, 1995.

[8] Bottou L., Stochastic gradient learning in neural networks, *NEURO NIMES'91*, 687-706, 1991.

[9] Bottou L., Online learning and stochastic approximations, in *Online Learning in Neural Networks*, David Saad (Ed), pp. 9-42, Cambridge University Press, 1999.

[10] Brezezniak Z. and T. Zastawniak, *Basic Stochastic Processes*, Springer-Verlag Berlin Heidelberg New York, 1998.

[11] Doob J.L., *Stochastic processes*, John Wiley and Sons, New York, 1953.

[12] Edwards P.J. and A.F. Murray, Can deterministic penalty terms model the effects of synaptic weight noise on network fault-tolerance? *International Journal of Neural Systems*, 6(4):401-16, 1995.

[13] Edwards P.J. and A.F. Murray, Fault tolerant via weight noise in analog VLSI implementations of MLP's – A case study with EPSILON, *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, Vol.45, No.9, p.1255-1262, Sep 1998.

[14] Ho K., C.S. Leung, and J. Sum, On weight-noise-injection training, M.Koeppen, N.Kasabov and G.Coghill (Eds.), *Advances in Neuro-Information Processing*, Springer LNCS 5507, p.919-926, 2009.

[15] Ho K., C.S. Leung and J. Sum, Convergence and Objective Functions of Some Fault/Noise Injection-Based On-line Learning Algorithms for RBF Networks, *IEEE Transactions on Neural Networks*, Vol.21(8), 1232-1244, August, 2010.

[16] Ho K., C.S. Leung and J. Sum, Objective functions of online weight noise injection training algorithms for MLP, in submission.

[17] Jim K.C., C.L. Giles and B.G. Horne, An analysis of noise in recurrent neural networks: Convergence and generalization, *IEEE Transactions on Neural Networks*, Vol.7, 1424-1438, 1996.

[18] Leung C.S., G.H. Young, J. Sum and W.K. Kan, On the regularization of forgetting recursive least square, *IEEE Transactions on Neural Networks*, Vol.10, 1842-1846, 1999.

[19] Murray A.F. and P.J. Edwards, Synaptic weight noise during multilayer perceptron training: fault tolerance and training improvements, *IEEE Transactions on Neural Networks*, Vol.4(4), 722-725, 1993.

[20] Murray A.F. and P.J. Edwards, Enhanced MLP performance and fault tolerance resulting from synaptic weight noise during training, *IEEE Transactions on Neural Networks*, Vol.5(5), 792-802, 1994.

[21] Sum J., C.S. Leung and K. Ho, On objective function, regularizer and prediction error of a learning algorithm for dealing with multiplicative weight noise, *IEEE Transactions on Neural Networks* Vol.20(1), pp.124-138, Jan, 2009.

[22] Sum J., K. Ho, SNIWD: Simultaneous weight noise injection with weight decay for MLP training, C.S. Leung and M. Lee and J.H. Chan (Eds), Neural Information Processing, Springer LNCS No.5863, p.494-501, 2009. Spring-Verlag Berlin Heidelberg.

[23] V. Tadic and S. Stankovic, Learning in neural networks by normalized stochastic gradient algorithm: Local convergence, in *Proc. 5th Seminar Neural Netw. Appl. Electr. Eng., Yugoslavia*, p.11V17, September 2000.

[24] Takase H., H. Kita and T. Hayashi, A study on the simple penalty term to the error function from the viewpoint of fault tolearnt training, *Proc. IJCNN 2004*, 1045-1050, 2004.

Appendix

The content of this appendix is adapted from Chapter 4, Theorem 4.2, in [10]. A stochastic process $\{\eta_t, t \geq 1\}$ is a supermartingale if $E[|\eta_t|] < \infty$ for all t and

$$E[\eta_{t+1}|\eta_t, \eta_{t-1}, \cdots, \eta_1] \leq \eta_t. \tag{84}$$

Lemma 5 (Doob's Martingale Convergence Theorem): If $\{\eta_t, t \geq 1\}$ is a supermartingale such that for some $\phi < \infty$ and $E[|\eta_t|] \leq \phi$ for all t, then $\lim_{t\to\infty} \eta_t$ exists and is finite with probability one.