Effect of Weight Noise and Gradient Noise on Learning

Chang Su[1], John Sum[2]

Institute of Technology Management, National Chung Hsing University
Taichung 40227, Taiwan
[1]kayone119@gmail.com
[2]pfsum@nchu.edu.tw

***Abstract*— In our recent works, we have analyzed the convergence properties and the objective functions of the multiplicative noise and additive noise injection-based algorithms respectively. In this paper, we generalize our work to gradient systems with multiplicative and additive noise injection algorithm. Besides, we analyze the effect of the noise on the learning algorithms completely. It is found that learning with the additive noise can improve the generalization ability of a neural network, while learning with the multiplicative noise is not. We also found that if $F(\mathbf{x})$ satisfies the Lipschitz condition, under mild conditions on the step size, with the probability one the weight vector converges to a local minimum of the objective function.**

***Keywords*— additive noise, multiplicative noise, Langevin noise, weight noise injection, convergence**

1. INTRODUCE

Research on the effect of noise on neural networks has been conducted for almost two decades. Some researchers investigated the effect of noise on multilayer perceptrons (MLP) [5, 9, 3, 4, 21], recurrent neural networks (RNN) [16, 19] and associative networks [10, 17]. Their primary focus was on the effect of noise on the performance of a neural network and how a neural network can be designed to alleviate such effect. Some researchers analyzed the effects of additive input noise (AIN) [6, 18, 22, 23], additive weight noise (AWN) [8] and chaotic noise (CN) [1, 2] on back-propagation learning. Their primary focus was on the objective functions and the convergence analyses of these noise injection-based learning algorithms.

In recent years, the effects of injecting additive weight noise and multiplicative weight noise (MWN) on the RBF and MLP learning have been investigated [14, 15, 12, 13, 7]. By deriving the objective functions for the RBF learning algorithm with injecting AWN or MWN, it is found that injecting AWN or MWN during RBF learning cannot improve the generalization ability of an RBF [14, 7]. Similarly, by deriving the objective functions for the MLP learning algorithm with injecting AWN or MWN, it is found that injecting AWN can improve the generalization ability of an MLP but injecting MWN during MLP learning might not [15, 13, 7]. These results clarify a common missconception that injecting noise during learning must be able to improve the generalization ability of a neural network.

In our previous work, we have been focus on the objective function and convergence behaviours of the AWN and MWN noise respectively. In this paper, we would like to investigate AWN and MWN simultaneously. We consider a general model for learning algorithms that are developed based on gradient descent. The weight vector is corrupted with AWN and MWN simultaneously. The gradient vector is corrupted with non-zero mean Langevin noise.

We denote that $\mathbf{x}(t) \in R^n$ are training inputs of an unknown system and $F(\mathbf{x}) \in R$ is the objective function. In the next section, the model of learning is introduced. The objective function is derived Section 3. With this objective function, the convergence analysis is presented in Section 3. Finally. Section 4

gives the conclusion of the paper.

2. MODEL

Let $\mathbf{x}(t) \in R^n$ and $F(\mathbf{x}) \in R$ is a bounded scalar function of $\mathbf{x}$. Besides, it is assumed that $F(\mathbf{x})$ is differentiable up to third order. The gradient system with forgetting is defined as follows :

$$\begin{aligned} & \mathbf{x}(t+1) \\ = \; & \mathbf{x}(t) \\ & -\mu_t \left(\frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}} + \lambda \mathbf{x} + (\alpha \mathbf{e} + \mathbf{b}_L(t)) \right). \end{aligned} \tag{1}$$

where $\alpha > 0$ and $\mu_t > 0$, $\mu_t \to 0$ is the step size at the t^{th} step and

$$\frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}} = \left. \frac{\partial F(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}(t)}.$$

With multiplicative and additive noise, the vector $\mathbf{x}(t)$ in (1) is replaced by $\tilde{\mathbf{x}}(t)$, where

$$\tilde{\mathbf{x}}(t) = \mathbf{x}(t) + \mathbf{b}_A(t) + \mathbf{b}_M(t) \otimes \mathbf{x}(t). \tag{2}$$

In (1) and (2), $\mathbf{b}_L(t), \mathbf{b}_A(t), \mathbf{b}_M(t) \in R^n$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $S_L \mathbf{I}_{n\times n}$, $S_A \mathbf{I}_{n\times n}$, $S_M \mathbf{I}_{n\times n}$ respectively. $\otimes$ is the elementwise multiplication operator, i.e.

$$\begin{aligned} \mathbf{b}_A(t) &= (b_{A1}(t), \cdots, b_{An}(t))^T \\ \mathbf{b}_M(t) \otimes \mathbf{x}(t) &= (b_{M1}(t)x_1(t), \cdots, b_{Mn}(t)x_n(t))^T \\ \mathbf{b}_L(t) &= (b_{L1}(t), \cdots, b_{Ln}(t))^T \end{aligned}$$

The model (1) is given as follows :

$$\begin{aligned} & \mathbf{x}(t+1) \\ = \; & \mathbf{x}(t) \\ & -\mu_t \left(\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}} + \lambda \tilde{\mathbf{x}}(t) + (\alpha \mathbf{e} + \mathbf{b}_L(t)) \right). \end{aligned} \tag{3}$$

Here, we assume that $E[b_{Li}(t)] = 0$ for all $i = 1, \cdots, n$ and $t \geq 0$. $E[b_{Li}^2(t)]$ equals to S_L and $E[b_{Li}(t)b_{Lj}(t)]$ equals zero if $i \neq j$. $E[b_{Li}(t_1)b_{Li}(t_2)] = 0$ if $t_1 \neq t_2$. These conditions applies to $\mathbf{b}_A(t)$ and $\mathbf{b}_M(t)$ as well.

In the rest of the paper, we make the following assumptions on the function $F(\cdot)$ and the noise.

- $F(\mathbf{x})$ satisfies the Lipschitz condition. There exists a constant K such that

$$|F(\mathbf{x}) - F(\mathbf{x}')| \leq K \|\mathbf{x} - \mathbf{x}'\|_2 \tag{4}$$

  for all $\mathbf{x}$ and $\mathbf{x}'$.

- S_A, S_M and S_L are small.

3. EFFECT OF NOISE

3.1. Objective Function

Given $\mathbf{x}(t)$, we get the mean update of (3) that

$$\begin{aligned} & E[\mathbf{x}(t+1)|\mathbf{x}(t)] \\ = \; & \mathbf{x}(t) \\ & -\mu_t E\left[\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}} + \lambda \tilde{\mathbf{x}}(t) \middle| \mathbf{x}(t) + (\alpha \mathbf{e} + \mathbf{b}_L(t)) \right]. \end{aligned} \tag{5}$$

In (5), the expectation is taken over the probability space of $\tilde{\mathbf{x}}(t)$. Since $E[\mathbf{b}_L(t)] = \mathbf{0}, E[\mathbf{b}_A(t)] = \mathbf{0}, E[\mathbf{b}_M(t)] = \mathbf{0}, E[\alpha] = \alpha$. Equation (5) can be rewritten as follows :

$$\begin{aligned} & E[\mathbf{x}(t+1)|\mathbf{x}(t)] \\ = \; & \mathbf{x}(t) - \mu_t \left(E\left[\frac{\partial F(\tilde{\mathbf{x}})}{\partial \mathbf{x}} \middle| \mathbf{x}(t) \right] + \lambda \mathbf{x}(t) + \alpha \mathbf{e} \right). \end{aligned} \tag{6}$$

Next, we let $V_{\otimes}(\mathbf{x})$ be a scalar function such that

$$E[\mathbf{x}(t+1)|\mathbf{x}(t)] = \mathbf{x}(t) - \mu_t \frac{\partial V_{\otimes}(\mathbf{x}(t))}{\partial \mathbf{x}}. \tag{7}$$

It can be shown the follow theorem.

Theorem 1 *For a gradient system defined as (1) and $\mathbf{x}(t)$ is corrupted by multiplicative and additive noise as stated in (2),*

$$\begin{aligned} E[F(\tilde{\mathbf{x}})|\mathbf{x}] &= F(\mathbf{x}) + \frac{S_A}{2} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} x_j^2 \\ &\quad + \frac{S_M}{2} \sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} x_j^2 \end{aligned} \tag{8}$$

and

$$\begin{aligned} V_{\otimes}(\mathbf{x}) &= F(\mathbf{x}) + \frac{S_A}{2}\sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} \\ &+ \frac{S_M}{2}\sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j^2}x_j^2 \\ &+ \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \alpha\sum_{j=1}^{n}x_j \\ &- S_M \int \mathbf{x} \otimes \mathbf{diag}\left\{\mathbf{H}(\mathbf{x})\right\} \cdot d\mathbf{x}. \quad (9) \end{aligned}$$

where $\int$ *is the line integral,* $\mathbf{H}(\mathbf{x})$ *is the Hessian matrix of* $F(\mathbf{x})$*, i.e.* $\mathbf{H}(\mathbf{x}) = \nabla\nabla_{\mathbf{x}}F(\mathbf{x})$ *and*

$$\mathbf{diag}\left\{\mathbf{H}(\mathbf{x})\right\} = \left(\frac{\partial^2 F(\mathbf{x})}{\partial x_1^2}, \frac{\partial^2 F(\mathbf{x})}{\partial x_2^2}, \cdots, \frac{\partial^2 F(\mathbf{x})}{\partial x_n^2}\right)^T.$$

Proof: Consider (6) and let $\frac{\partial F(\mathbf{x})}{\partial x_i}$ be the i^{th} element of $\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}$.

$$\begin{aligned} &\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i} \\ &= \frac{\partial F(\mathbf{x})}{\partial x_i} + \sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i}(b_{Aj} + b_{Mj}x_j) \\ &+ \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_k \partial x_j \partial x_i}b_{Ak}b_{Aj} \\ &+ \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_k \partial x_j \partial x_i}b_{Ak}b_{Mj}x_j \\ &+ \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_k \partial x_j \partial x_i}b_{Aj}b_{Mk}x_k \\ &+ \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_k \partial x_j \partial x_i}b_{Mk}b_{Mj}x_k x_j. \end{aligned} \quad (10)$$

Therefore,

$$\begin{aligned} E\left[\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i}\middle|\mathbf{x}\right] &= \frac{\partial F(\mathbf{x})}{\partial x_i} \\ &+ \frac{S_A}{2}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_j \partial x_j \partial x_i} \\ &+ \frac{S_M}{2}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_j \partial x_j \partial x_i}x_j^2. \end{aligned} \quad (11)$$

On the other hand,

$$\begin{aligned} F(\tilde{\mathbf{x}}) &= F(\mathbf{x}) + \sum_{i=1}^{n}\frac{\partial F(\mathbf{x})}{\partial x_i}(b_{Ai} + b_{Mi}x_i) \\ &+ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i}b_{Ak}b_{Aj} \\ &+ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i}b_{Ak}b_{Mj}x_j \\ &+ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i}b_{Aj}b_{Mk}x_k \\ &+ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i}b_{Mk}b_{Mj}x_k x_j). \end{aligned} \quad (12)$$

Thus,

$$\begin{aligned} E[F(\tilde{\mathbf{x}})|\mathbf{x}] &= F(\mathbf{x}) + \frac{S_A}{2}\sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} \\ &+ \frac{S_M}{2}\sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j}x_j^2, \end{aligned} \quad (13)$$

and

$$\begin{aligned} \frac{\partial}{\partial x_i}E[F(\tilde{\mathbf{x}})|\mathbf{x}] &= \frac{\partial F(\mathbf{x})}{\partial x_i} + \frac{S_A}{2}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_i \partial x_j \partial x_j} \\ &+ \frac{S_M}{2}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_i \partial x_j \partial x_j}x_j^2 \\ &+ S_M\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_i}x_i. \end{aligned} \quad (14)$$

By the fact that

$$\frac{\partial^3 F(\mathbf{x})}{\partial x_j \partial x_j \partial x_i} = \frac{\partial^3 F(\mathbf{x})}{\partial x_i \partial x_j \partial x_j} \quad (15)$$

for $F(\mathbf{x})$ is triple differentiable. Compare (11) and (14), we get that

$$E\left[\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i}\middle|\mathbf{x}\right] = \frac{\partial}{\partial x_i}E[F(\tilde{\mathbf{x}})|\mathbf{x}] - S_M\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_i}x_i. \quad (16)$$

Further by (6) and (7), we get that

$$\frac{\partial V_{\otimes}(\mathbf{x})}{\partial x_i} = \frac{\partial}{\partial x_i}E[F(\tilde{\mathbf{x}})|\mathbf{x}] - S_M\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_i} + \lambda x_i(t) + \alpha. \quad (17)$$

$$
\begin{aligned}
V_{\otimes}(\mathbf{x}) &= E[F(\tilde{\mathbf{x}})|\mathbf{x}] \\
&\quad -S_M \int \mathbf{x} \otimes \mathbf{diag}\left\{\mathbf{H}(\mathbf{x})\right\} \cdot d\mathbf{x} \\
&\quad +\frac{\lambda}{2}\|x\|_2^2 + \alpha \sum_{j=1}^{n} x_j. \qquad (18)
\end{aligned}
$$

In other word,

$$
\begin{aligned}
V_{\otimes}(\mathbf{x}) &= F(\mathbf{x}) + \frac{S_A}{2} \sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} \\
&\quad +\frac{S_M}{2} \sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j^2} x_j^2 \\
&\quad +\frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \alpha \sum_{j=1}^{n} x_j \\
&\quad -S_M \int \mathbf{x} \otimes \mathbf{diag}\left\{\mathbf{H}(\mathbf{x})\right\} \cdot d\mathbf{x}.
\end{aligned} \qquad (19)
$$

Then, the proof is completed. **Q.E.D.**

3.2. Convergence

The effect of the first four terms are to bring $\mathbf{x}$ closer to the zero vector while the last term is to push it away from the zero vector. Therefore, the existence of multiplicative and additive noise in a gradient system would lead to both regularization effect and de-regularization effect.

Now, we proceed to the convergence analysis. The proof is conducted by the following steps. First, we show that $E[\|\mathbf{x}(t)\|_2]$ is bounded for all $t \geq 0$ and $\lim_{t\to\infty} \|\mathbf{x}(t)\|_2$ exists. It implies that for sufficient large t, the elements in $\mathbf{x}(t)$ must be bounded. Then apply these two results to show the convergence of the system (3).

Consider (3), let

$$
\eta = (1 - \mu_t\lambda(1 - S_M))\mathbf{x}(t) - \mu_t \frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}}
$$

we get that

$$
\begin{aligned}
& E[\|\mathbf{x}(t+1)\|_2^2|\mathbf{x}(t)] \\
= \;& E\left[\left\|(1 - \mu_t\lambda)\mathbf{x}(t) - \mu_t \frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}}\right\|_2^2 \middle| \mathbf{x}(t)\right] \\
& +\mu_t^2 \left(\lambda^2(S_A + S_M\|\mathbf{x}(t)\|_2^2) + \alpha^2 + nS_L\right), \\
\leq \;& E\left[\|\eta\|_2^2 \,|\mathbf{x}(t)\right] + \mu_t^2 \left(\lambda^2 S_A + \alpha^2 + nS_L\right). \qquad (20)
\end{aligned}
$$

By Jensen inequality [20], we get that

$$
E[\|\mathbf{x}(t+1)\|_2|\mathbf{x}(t)] \leq \left(E[\|\mathbf{x}(t+1)\|_2^2|\mathbf{x}(t)]\right)^{1/2}. \qquad (21)
$$

Therefore, by (20) and (21), we get that

$$
\begin{aligned}
& E[\|\mathbf{x}(t+1)\|_2|\mathbf{x}(t)] \\
\leq \;& E\left[\|\eta\|_2 \,|\mathbf{x}(t)\right] + \mu_t\sqrt{\lambda^2 S_A + \alpha + nS_L} \\
\leq \;& (1 - \mu_t\lambda(1 - S_M))\|\mathbf{x}(t)\|_2 \\
& +\mu_t E\left[\left\|\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}}\right\|_2 \middle| \mathbf{x}(t)\right] \\
& +\mu_t\sqrt{\lambda^2 S_A + \alpha + nS_L}.
\end{aligned} \qquad (22)
$$

The last inequality is due to Triangle inequality. Note that

$$
\frac{F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})}{\Delta x_i} \leq \frac{|F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})|}{|\Delta x_i|},
$$

where $\Delta\mathbf{x} = (x_1, \cdots, x_{i-1}, x_i + \delta x_i, x_{i+1}, \cdots, x_n)^T$. Recall that $F(\mathbf{x})$ satisfies the Lipschitz condition (4). By the Lipschitz condition, we can get that

$$
\begin{aligned}
& E\left[\left\|\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}}\right\|_2 \middle| \mathbf{x}(t)\right] \\
= \;& \sqrt{\left(\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}_1}\right)^2 + \cdots + \left(\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}_n}\right)^2} \\
\leq \;& \sqrt{n}K.
\end{aligned}
$$

As a result, we can get by (22) that

$$
E[\|\mathbf{x}(t+1)\|_2|\mathbf{x}(t)] \leq (1 - \mu_t\gamma)\|\mathbf{x}(t)\|_2 + \mu_t\kappa_1, \qquad (23)
$$

where

$$
\begin{aligned}
\gamma &= \lambda(1 - S_M), \\
\kappa_1 &= \sqrt{\lambda^2 S_A + \alpha + nS_L} + \sqrt{n}K.
\end{aligned}
$$

Hence,

$$
E[\|\mathbf{x}(t+1)\|_2] \leq (1-\mu_t\gamma)E[\|\mathbf{x}(t)\|_2] + \mu_t\kappa_1. \qquad (24)
$$

Lemma 1 *If $0 < \mu_t\gamma < 1$ for all $t \geq 0$, $E[\|\mathbf{x}(t)\|_2]$ is bounded.*

Proof: Let $\kappa_2 = \kappa_1/\gamma$, we get the update of (24)

$$E[\|\mathbf{x}(t+1)\|_2] \leq (1-\mu_t\gamma)E[\|\mathbf{x}(t)\|_2] + \mu_t\gamma\kappa_2.$$

Thus,

$$\begin{aligned} & E[\|\mathbf{x}(t+1)\|_2] - \kappa_2 \\ \leq\ & (1-\mu_t\gamma)(E[\|\mathbf{x}(t)\|_2] - \kappa_2) \\ \leq\ & \prod_{\tau=0}^{t}(1-\mu_\tau\gamma)(E[\|\mathbf{x}(t)\|_2] - \kappa_2). \end{aligned} \tag{25}$$

Since $E[\|\mathbf{x}(0)\|_2]$ is bounded. $0 < \mu_t\gamma < 1$ implies that

$$0 < \prod_{\tau=0}^{t}(1-\mu_\tau\gamma) < 1.$$

Therefore,

$$\begin{aligned} E[\|\mathbf{x}(t+1)\|_2] - \kappa_2 &< E[\|\mathbf{x}(0)\|_2] - \kappa_2, \\ E[\|\mathbf{x}(t+1)\|_2] &< E[\|\mathbf{x}(0)\|_2]. \end{aligned} \tag{26}$$

$E[\|\mathbf{x}(t)\|_2]$ is bounded for all $t \geq 0$. The proof is completed. **Q.E.D.**

Lemma 2 *With probability one,* $\lim_{t\to\infty}\|\mathbf{x}(t)\|_2$ *exists.*

Proof: We can define a random variable as follows :

$$\begin{aligned} \beta(t) &= \|\mathbf{x}(t)\|_2\prod_{\tau=t}^{\infty}(1-\mu_t\gamma) \\ &\quad +\kappa_1\sum_{\tau_1=t}^{\infty}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{\infty}(1-\mu_{\tau_2}\gamma). \end{aligned} \tag{27}$$

By Lemma 1, $E[\|\mathbf{x}(t)\|_2]$ is bounded. It is clear that $\beta(t) \geq 0$. Now, we show that $\beta(t)$ is supermartingale. The expectation of $\beta(t+1) \geq 0$ is given by

$$\begin{aligned} & E[\beta(t+1)|\beta(t)] \\ =\ & E[\|\mathbf{x}(t+1)\|_2|\beta(t)]\prod_{\tau=t+1}^{\infty}(1-\mu_t\gamma) \\ & +\kappa_1\sum_{\tau_1=t+1}^{\infty}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{\infty}(1-\mu_{\tau_2}\gamma). \end{aligned} \tag{28}$$

By (23), we can get that

$$\begin{aligned} & E[\beta(t+1)|\beta(t)] \\ \leq\ & (1-\mu_t\gamma)\|\mathbf{x}(t)\|_2\prod_{\tau=t+1}^{\infty}(1-\mu_t\gamma) \\ & +\mu_t\kappa_1\prod_{\tau=t+1}^{\infty}(1-\mu_t\gamma) \\ & +\kappa_1\sum_{\tau_1=t+1}^{\infty}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{\infty}(1-\mu_{\tau_2}\gamma) \\ =\ & \|\mathbf{x}(t)\|_2\prod_{\tau=t+1}^{\infty}(1-\mu_t\gamma) \\ & +\kappa_1\sum_{\tau_1=t}^{\infty}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{\infty}(1-\mu_{\tau_2}\gamma). \end{aligned} \tag{29}$$

Clearly, the RHS of (29) is equal to $\beta(t)$. Thus, $E[\beta(t+1)|\beta(t)] \leq \beta(t)$ and

$$E[\beta(t+1)] \leq E[\beta(t) \leq \cdots \leq E[\beta(0)]. \tag{30}$$

Next, we are going to show that $E[\beta(0)]$ is finite. As $\|mathbfx(0)\|$ is finite, what we will show is that the second term in the RHS of (29) is finite. Let

$$\xi_T = \exp\left\{-\gamma\sum_{\tau_2=0}^{T}\mu_{\tau_2}\right\}.$$

By the inequality that

$$\ln(1-\mu_t\gamma) \leq -\mu_t\gamma,$$

$$
\begin{aligned}
&\sum_{\tau_1=0}^{T-1}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{T}(1-\mu_{\tau_2}\gamma)\\
\leq\ &\sum_{\tau_1=0}^{T-1}\mu_{\tau_1}\exp\left\{-\gamma\sum_{\tau_2=\tau 1+1}^{T}\mu_{\tau_2}\right\}\\
=\ &\xi_T\sum_{\tau_1=0}^{T-1}\mu_{\tau_1}\exp\left\{-\gamma\sum_{\tau_2=0}^{\tau_1}\mu_{\tau_2}\right\}\\
=\ &\xi_T\mu_0\exp\{\gamma\mu_0\}\\
+\ &\xi_T\sum_{\tau_1=1}^{T-1}\mu_{\tau_1}\exp\{\gamma\mu_{\tau_1}\}\exp\left\{\gamma\sum_{\tau_2=0}^{\tau_1-1}\mu_{\tau_2}\right\}\\
\leq\ &\xi_T\mu_0\exp\{\gamma\mu_0\}\\
+\ &\xi_T\exp\left\{\gamma\max_t\mu_t\right\}\sum_{\tau_1=1}^{T-1}\mu_{\tau_1}\exp\left\{\gamma\sum_{\tau_2=0}^{\tau_1-1}\mu_{\tau_2}\right\}
\end{aligned}
\tag{31}
$$

and

$$
\begin{aligned}
&\sum_{\tau_1=1}^{T-1}\mu_{\tau_1}\exp\left\{\gamma\sum_{\tau_2=0}^{\tau_1-1}\mu_{\tau_2}\right\}\\
\leq\ &\int_{\mu_0}^{\bar{T}-\mu_T}\exp(\gamma x)dx\\
<\ &\int_0^{\bar{T}}\exp(\gamma x)dx.
\end{aligned}
\tag{32}
$$

where $\bar{T}=\sum_{\tau=0}^{T}\mu_\tau$. From (31) and (32),

$$
\begin{aligned}
&\sum_{\tau_1=0}^{T-1}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{T}(1-\mu_{\tau_2}\gamma)\\
<\ &\xi_T\mu_0\exp\gamma\mu_0\\
&+\frac{\exp\{\gamma\max_t\mu_t\}}{\gamma}\left\{1-\exp(-\gamma\bar{T})\right\}.
\end{aligned}
\tag{33}
$$

Since $\lim_{T\to\infty}\xi_T=0$,

$$
\begin{aligned}
&\lim_{T\to\infty}\sum_{\tau_1=0}^{T-1}\mu_{\tau_1}\prod_{\tau_2=\tau_1+1}^{T}(1-\mu_{\tau_2}\gamma)\\
\leq\ &\frac{\exp\{\gamma\max_t\mu_t\}}{\gamma}.
\end{aligned}
\tag{34}
$$

Finally, from (27) and the fact that $\|\mathbf{x}(0)\|_2$ is finite, we get

$$
E[\beta(0)]\leq\kappa_1\frac{\exp\{\gamma\max_t\mu_t\}}{\gamma}
$$

and thus

$$
E[\beta(t+1)]\leq E[\beta(t)\leq\cdots\leq E[\beta(0)]<\infty. \tag{35}
$$

By Martingale Convergence Theorem, $\lim_{t\to\infty}\beta(t)$ exists with probability one. By (34), it is clear that the value of the second term in (27) is bounded and positive. On the other hand, the factor $\prod_{\tau=t}^{\infty}(1-\mu_t\gamma)$ associated with $\|\mathbf{x}\|_2$ in (27) is increasing with respect to t. Its value is positive and bounded by one. Therefore, we can conclude that $\lim_{t\to\infty}\|\mathbf{x}\|_2$ exists with probability one. The proof is completed.

As $\lim_{t\to\infty}\|\mathbf{x}\|_2$ exists, there exists a bounded region Ω and t^* such that $\mathbf{x}(t)\in\Omega$ for all $t\geq t^*$. Now, we can state the convergence of (3) in the following lemma and theorem.

Lemma 3 *For all $t\geq t^*$, there exists a bounded region Ω and t^* such that $\mathbf{x}(t)\in\Omega$; $V_\otimes(\mathbf{x}(t))\leq\infty$; and the eigenvalues of the Hessian matrix $\nabla\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))$ are all finite.*

Theorem 2 (Convergence) *For the learning algorithm defined as (1) and $\tilde{\mathbf{x}}(t)$ is defined as (2), if $\mu_t\to 0,\sum_t\mu_t=\infty$ and $\sum_t\mu_t^2<\infty$,then with probability one $\lim_{t\to\infty}\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))=\mathbf{0}$, where $\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))$ is given by (19).*

Proof: Now, we expand $V_\otimes(\mathbf{x}(t+1))$ around $\mathbf{x}(t)$ and get that

$$
\begin{aligned}
V_\otimes(\mathbf{x}(t+1))\ =\ &V_\otimes(\mathbf{x}(t))+\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))\delta\mathbf{x}(t)\\
&+\frac{1}{2}\delta\mathbf{x}(t)^T\nabla\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))\delta\mathbf{x}(t),
\end{aligned}
\tag{36}
$$

where $\delta\mathbf{x}(t)=\mathbf{x}(t+1)-\mathbf{x}(t)$. By Lemma 3, we can let κ_3 be the maximum eigenvalue of $\nabla\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))$ for $t\geq t^*$. By (36), we can get the inequality that

$$
\begin{aligned}
V_\otimes(\mathbf{x}(t+1))\ \leq\ &V_\otimes(\mathbf{x}(t))+\nabla_{\mathbf{x}}V_\otimes(\mathbf{x}(t))\delta\mathbf{x}(t)\\
&+\frac{\kappa_3}{2}\|\delta\mathbf{x}(t)\|_2^2.
\end{aligned}
\tag{37}
$$

By Lemma 3, it is clear that $\|\delta\mathbf{x}(t)\|_2^2$ is bounded for all $t \geq t^*$. We let this bound be κ_4. As a result, we can get that

$$\begin{aligned} & \lim_{t\to\infty} E[V_{\otimes}(\mathbf{x}(t))|\mathbf{x}(t^*)] \\ \leq \; & V_{\otimes}(\mathbf{x}(t^*)) - \sum_{t\geq t^*} \mu_t E[\|\nabla_{\mathbf{x}} V_{\otimes}(\mathbf{x}(t))\|_2^2 \,|\mathbf{x}(t^*)] \\ & + \frac{\kappa_3\kappa_4}{2} \sum_{t\geq t^*} \mu_t^2. \end{aligned} \tag{38}$$

It is clear from Lemma 3 that $\lim_{t\to\infty} E[V_{\otimes}(\mathbf{x}(t))|\mathbf{x}(t^*)]$ and $V_{\otimes}(\mathbf{x}(t^*))$ are all finite. Further from the condition that $\sum_{t\geq t^*} \mu_t^2 < \infty$, we can get that

$$\sum_{t\geq t^*} \mu_t E\left[\|\nabla_{\mathbf{x}} V_{\otimes}(\mathbf{x}(t))\|_2^2 \,|\mathbf{x}(t^*)\right] < \infty.$$

By the condition that $\sum_{t\geq t^*} \mu_t = \infty$, we can prove by contradiction that

$$\lim_{t\to\infty} E\left[\|\nabla_{\mathbf{x}} V_{\otimes}(\mathbf{x}(t))\|_2^2 \,|\mathbf{x}(t^*)\right] = 0$$

In the other word,

$$\lim_{t\to\infty} \nabla_{\mathbf{x}} V_{\otimes}(\mathbf{x}(t)) = \mathbf{0}.$$

The proof is completed.

4. CONCLUSION

In this paper, we have presented the objective functions and convergence analyses of multiplicative and additive noise on learning. The gradient vector is corrupted with non-zero mean Langevin noise. The energy functions of the gradient system with noise has been released. We also show that with probability one the weight vector converge to a local minimum of the objective function. Our result imply that, with the multiplicative and additive noise on learning, two opposite effects exists, moving towards and away. The result shows that inject AWN and MWN simultaneously might not be improve generalization.

ACKNOWLEDGEMENT

The work presented in this paper is supported in part by research grants from Taiwan National Science Council numbering 100-2221-E-126-015 and 101-2221-E-126-016.

REFERENCES

[1] A. Azamimi, Y. Uwate, and Y. Nishio, *An analysis of chaotic noise injected to backpropagation algorithm in feedforward neural network*, Proceedings of IWVCC08, 70-73, 2008.

[2] A. Azamimi, Y. Uwate, and Y. Nishio, *Effect of chaos noise on the learning ability of back propagation algorithm in feed forward neural network*, Proceedings of the 6th Internationa Colloquium on Signal Processing and Its Applications (CSPA), 2010.

[3] A.F. Murray 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.

[4] A.F. Murray 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.

[5] C.H. Sequin and R.D. Clay, *Fault tolerance in feedforward artificial neural networks*, Neural Networks, Vol.4, 111-141, 1991.

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

[7] C. Su, J. Sum, C. S. Leung, K. I.-J. Ho, *Noise on gradient systems with forgetting*, in S. Frik et al. (Eds): ICONIP2015, PartIII, LNCS 9491, 98. 1-9, 2015.

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

[9] G. Bolt, *Fault tolerant in multi-layer Perceptrons*. PhD Thesis, University of York, UK, 1992.

[10] H. Asai, K. Onodera, T. Kamio, H. Ninomiya, *A study of Hopfield neural networks with external noises*, Proceedings IEEE International Conference on Neural Networks, Vol. 4, 1584-1589, 1995.

[11] H. Zhang, Y. Zhang, D. Xu, X. Liu, *Deterministic convergence of chaos injection-based gradient method for training feedforward neural networks*, to appear in Cognitive Neurodynamic.

[12] J. Sum, C.S. Leung, K. Ho, *Convergence analysis of on-line node fault injection-based training algorithms for MLP networks*, IEEE Transactions on Neural Networks and Learning Systems, Vol.23(2), 211-222, Feb 2012.

[13] J. Sum, C.S. Leung, K. Ho, *Convergence analyses on on-line weight noise injection-based training algorithms for MLPs*, IEEE Transactions on Neural Networks and Learning Systems, Vol.23(11), 1827-1840, Nov 2012.

[14] K. Ho, C.S. Leung and J. Sum, *Convergence and objective functions of Some Fault/Noise Injection-Based online Learning Algorithms for RBF Networks*, IEEE Transactions on Neural Networks, Vol.21(6), 938-947, June, 2010.

[15] K. Ho, C.S. Leung, J. Sum, *Objective functions of the online weight noise injection training algorithms for MLP*, IEEE Transactions on Neural Networks, Vol.22(2), 317-323, Feb 2011.

[16] K.C. Jim, 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.

[17] L. Wang, *Noise injection into inputs in sparsely connected Hopfield and winner-take-all neural networks*, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, Vol.27(5), 868-870, October 1997.

[18] R. Reed, R.J. Marks II & S. Oh, *Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter*, IEEE Transactions on Neural Networks, Vol.6(3), 529-538, 1995.

[19] S. Das and O. Olurotimi, *Noisy recurrent neural networks: The continuous-time case*, IEEE Transactions on Neural Networks, Vol.9(5), 913-935, 1998.

[20] S.M. Ross, Stochastic Process. New York: Wiley, 1996.

[21] T. Rögnvaldsson, *On Langevin updating in multilayer perceptrons*, Neural Computation, Vol.6(5), 916-926, 1994.

[22] Y. Grandvalet, S. Canu, *A comment on noise injection into inputs in back-propagation learning*, IEEE Transactions on Systems, Man, and Cybernetics, 1995.

[23] Y. Grandvalet, S. Canu, S. Boucheron, *Noise injection : Theoretical prospects*, Neural Computation, 1997.