Noise on Gradient Systems with Forgetting

Chang Su[1], John Sum[1](✉), Chi-Sing Leung[2], and Kevin I.-J. Ho[3]

[1] ITM, National Chung Hsing University, Taichung, Taiwan
pfsum@nchu.edu.tw
[2] EE, City University of Hong Kong, Kowloon Tong, Hong Kong
eeleungc@cityu.edu.hk
[3] CSCE, Providence University, Taichung, Taiwan
ho@pu.edu.tw

Abstract. In this paper, we study the effect of noise on a gradient system with forgetting. The noise include multiplicative noise, additive noise and chaotic noise. For multiplicative or additive noise, the noise is a mean zero Gaussian noise. It is added to the state vector of the system. For chaotic noise, it is added to the gradient vector. Let $\mathbf{x}$ be the state vector of a system, S_b be the variance of the Gaussian noise, κ' is average noise level of the chaotic noise, λ is a positive constant, $V(\mathbf{x})$ be the energy function of the original gradient system, $V_{\otimes}(\mathbf{x})$, $V_{\oplus}(\mathbf{x})$ and $V_{\odot}(\mathbf{x})$ be the energy functions of the gradient systems, if multiplicative, additive and chaotic noises are introduced. Suppose $V(\mathbf{x}) = F(\mathbf{x}) + \lambda\|\mathbf{x}\|_2^2$. It is shown that $V_{\otimes}(\mathbf{x}) = V(\mathbf{x}) + (S_b/2)\sum_{j=1}^{n}(\partial^2 F(\mathbf{x})/\partial x_j^2)x_j^2 - S_b\sum_{j=1}^{n}\int x_j(\partial^2 F(\mathbf{x})/\partial x_j^2)dx_j$, $V_{\oplus}(\mathbf{x}) = V(\mathbf{x}) + (S_b/2)\sum_{j=1}^{n}\partial^2 F(\mathbf{x})/\partial x_j^2$, and $V_{\odot}(\mathbf{x}) = V(\mathbf{x}) + \kappa'\sum_{i=1}^{n} x_i$. The first two results imply that multiplicative or additive noise has no effect on the system if $F(\mathbf{x})$ is quadratic. While the third result implies that adding chaotic noise can have no effect on the system if κ' is zero. As many learning algorithms are developed based on the method of gradient descent, these results can be applied in analyzing the effect of noise on those algorithms.

1 Introduction

Research on the effect of noise on neural networks has been conducted for almost two decades. From the earlier 90 s to the mid 90 s, researchers investigated the effect of noise on the performance of a multilayer perceptron (MLP)/recurrent neural networks (RNN) [6,11–13,15] and the associative networks [2,19]. Later, from the mid 90 s to the late 90 s, researchers started to analyze the effects of *additive input noise* (AIN) [5,7,8,14] and additive weight noise (AWN) [1] on back-propagation learning. The objective functions for these noise injection-based learning algorithms were revealed. In the 2000s, researchers investigated the effect of *chaotic noise* (CN) on MLP [3,4].

In recent years, the effects of AWN and *multiplicative weight noise* (MWN) on the RBF and MLP learning algorithms have been investigated [9,10,16,17]. It is shown that the objective function of the RBF learning algorithm with adding AWN or MWN is identical to the original RBF learning algorithm [9].

S. Arik et al. (Eds.): ICONIP 2015, Part III, LNCS 9491, pp. 479–487, 2015.
DOI: 10.1007/978-3-319-26555-1_54

Hence, adding AWN or MWN during RBF learning cannot improve the generalization ability of an RBF. Adding AWN during MLP learning can improve the generalization ability of an MLP. Adding MWN during MLP learning might not be [10,17]. These results clarify a common missconception that adding noise during learning is able to improve the generalization ability of a neural network.

Now, we would like to investigate another question. *Would similar results be obtained for other learning algorithms?* To do so, one obvious approach is to investigate the effect of noise on a gradient system as many learning algorithms are developed by the method of gradient descent. Understanding the effect of noise on gradient systems can aid in the understanding of the effect of noise on these learning algorithms. Therefore, the objective of the paper is to investigate the effects of three types of noise (multiplicative noise, additive noise and chaotic noise) on a gradient system with forgetting. The energy functions of the corresponding gradient systems are revealed.

In the next section, the gradient systems with noise are introduced. The energy functions of these gradient systems with noise will be analyzed in Sect. 3. Effect of noise on the gradient systems will be elucidated in Sect. 4. Finally. Section 5 gives the conclusion of the paper.

2 Models

Let $\mathbf{x}(t) \in R^n$ and $F(\mathbf{x}) \in R$ is a bounded smooth function of $\mathbf{x}$. The energy function is given by $V(\mathbf{x}) = F(\mathbf{x}) + \lambda\|\mathbf{x}\|_2^2$, where λ is a small positive number called forgetting factor. The gradient system is defined as follows:

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \mu\left(\frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}} + \lambda\mathbf{x}(t)\right), \tag{1}$$

where μ is the learning step and it is a small positive number, and $\partial F(\mathbf{x}(t))/\partial\mathbf{x} = \partial F(\mathbf{x})/\partial\mathbf{x}|_{\mathbf{x}=\mathbf{x}(t)}$.

2.1 Multiplicative/Additive Noise

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

$$\begin{aligned} \tilde{\mathbf{x}}(t) &= \mathbf{x}(t) + \mathbf{b}(t) \otimes \mathbf{x}(t). \\ \mathbf{b}(t) \otimes \mathbf{x}(t) &= (b_1(t)x_1(t), b_2(t)x_2(t), \cdots, b_n(t)x_n(t))^T. \end{aligned} \tag{2}$$

With additive noise,

$$\tilde{\mathbf{x}}(t) = \mathbf{x}(t) + \mathbf{b}(t). \tag{3}$$

In (2) and (3), $\mathbf{b}(t) \in R^n$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $S_b\mathbf{I}_{n\times n}$. Moreover, $E[b_i(t)] = 0$ for all $i = 1, \cdots, n$ and $t \geq 0$. $E[b_i^2(t)]$ equals to S_b and $E[b_i(t)b_j(t)]$ equals zero if $i \neq j$. $E[b_i(t_1)b_i(t_2)] = 0$ if $t_1 \neq t_2$. The gradient system with noise is given as follows:

$$\mathbf{x}(t+1) = \tilde{\mathbf{x}}(t) - \mu\left(\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}} + \lambda\tilde{\mathbf{x}}(t)\right). \tag{4}$$

2.2 Chaotic Noise

With chaotic noise injection, the noise is added to the gradient vector as follows [3,4,20]:

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \mu\left(\frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}} + \lambda \mathbf{x}(t) + \kappa n(t)\mathbf{e}\right), \tag{5}$$

where $\mathbf{e}$ is a constant vector of all 1s, κ is a positive constant and $n(t)$ is a deterministic noise generated by

$$n(t+1) = \alpha n(t)(1 - n(t)), \quad 3.6 < \alpha < 4. \tag{6}$$

3 Energy Functions

In this section, the energy functions of these gradient systems with noise are revealed. The effect of noise on the gradient systems will be discussed in the next section.

3.1 Multiplicative/Additive Noise

Given $\mathbf{x}(t)$, the mean update of (4) can be written as follows:

$$E[\mathbf{x}(t+1)|\mathbf{x}(t)] = E[\tilde{\mathbf{x}}(t)|\mathbf{x}(t)] - \mu E\left[\frac{\partial F(\tilde{\mathbf{x}}(t))}{\partial \mathbf{x}} + \lambda \tilde{\mathbf{x}}(t)\middle| \mathbf{x}(t)\right]. \tag{7}$$

In (7), the expectation is taken over the probability space of $\tilde{\mathbf{x}}(t)$. Since $E[\mathbf{b}(t)] = \mathbf{0}$, by (2) we get that $E[\tilde{\mathbf{x}}(t)|\mathbf{x}(t)] = \mathbf{x}(t)$. Equation (7) can be rewritten as follows:

$$E[\mathbf{x}(t+1)|\mathbf{x}(t)] = \mathbf{x}(t) - \mu\left(E\left[\frac{\partial F(\tilde{\mathbf{x}})}{\partial \mathbf{x}}\middle| \mathbf{x}(t)\right] + \lambda \mathbf{x}(t)\right). \tag{8}$$

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

$$E[\mathbf{x}(t+1)|\mathbf{x}(t)] = \mathbf{x}(t) - \mu \frac{\partial V_{\otimes}(\mathbf{x}(t))}{\partial \mathbf{x}}. \tag{9}$$

The energy function is stated in the following theorem.

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

$$E[F(\tilde{\mathbf{x}})|\mathbf{x}] = F(\mathbf{x}) + \frac{S_b}{2}\sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} x_j^2 \tag{10}$$

and

$$V_{\otimes}(\mathbf{x}) = F(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \frac{S_b}{2}\sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} x_j^2 - S_b \int \mathbf{x} \otimes \mathbf{diag}\{\mathbf{H}(\mathbf{x})\} \cdot d\mathbf{x}. \tag{11}$$

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 (8) and let $\partial F(\mathbf{x})/\partial x_i$ be the i^{th} element of $\partial F(\mathbf{x})/\partial \mathbf{x}$.

$$\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i} = \frac{\partial F(\mathbf{x})}{\partial x_i} + \sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i}(b_j x_j) + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial^3 F(\mathbf{x})}{\partial x_k \partial x_j \partial x_i} b_k b_j x_k x_j. \quad (12)$$

Therefore,

$$E\left[\left.\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i}\right| \mathbf{x}\right] = \frac{\partial F(\mathbf{x})}{\partial x_i} + \frac{S_b}{2}\sum_{j=1}^{n} \frac{\partial^3 F(\mathbf{x})}{\partial x_j \partial x_j \partial x_i} x_j^2. \quad (13)$$

On the other hand, by expanding $F(\tilde{\mathbf{x}})$ about $\mathbf{x}$, we get that

$$F(\tilde{\mathbf{x}}) = F(\mathbf{x}) + \sum_{i=1}^{n} \frac{\partial F(\mathbf{x})}{\partial x_i} b_i x_i + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i} b_j b_i x_j x_i$$

and hence

$$E[F(\tilde{\mathbf{x}})|\mathbf{x}] = F(\mathbf{x}) + \frac{S_b}{2}\sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} x_j^2. \quad (14)$$

Differentiate both side of (14) with respect to x_i, we get that

$$\frac{\partial}{\partial x_i} E[F(\tilde{\mathbf{x}})|\mathbf{x}] = \frac{\partial F(\mathbf{x})}{\partial x_i} + \frac{S_b}{2}\sum_{j=1}^{n} \frac{\partial^3 F(\mathbf{x})}{\partial x_i \partial x_j \partial x_j} x_j^2 + S_b \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_i} x_i. \quad (15)$$

As $F(\mathbf{x})$ is smooth, $\partial^3 F(\mathbf{x})/\partial x_j \partial x_j \partial x_i = \partial^3 F(\mathbf{x})/\partial x_i \partial x_j \partial x_j$. Compare (13) and (15), we get that

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

Further by (8) and (9), we get that

$$V_{\otimes}(\mathbf{x}) = E[F(\tilde{\mathbf{x}})|\mathbf{x}] - S_b \int \mathbf{x} \otimes \mathbf{diag}\left\{\mathbf{H}(\mathbf{x})\right\} \cdot d\mathbf{x} + \frac{\lambda}{2}\|x\|_2^2. \quad (16)$$

Putting (10) in (16) and rearranging the terms, we can get the energy function as given in (11) and the proof is completed. **Q.E.D.**

Similarly, the energy function of the gradient system with additive noise is stated in the following theorem.

Theorem 2. *For a gradient system defined as (1) and* $\mathbf{x}(t)$ *is corrupted by additive noise as stated in (3),*

$$V_{\oplus}(\mathbf{x}) = F(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j}. \tag{17}$$

Proof: For additive noise, the noisy $\tilde{\mathbf{x}}$ in (4) is given by $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{b}$. Similarly, we consider (8) and let $\partial F(\mathbf{x})/\partial x_i$ be the i^{th} element of $\partial F(\mathbf{x})/\partial \mathbf{x}$.

$$\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i} = \frac{\partial F(\mathbf{x})}{\partial x_i} + \sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_i} b_j + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_k \partial x_j \partial x_i} b_k b_j.$$

Therefore,

$$E\left[\frac{\partial F(\tilde{\mathbf{x}})}{\partial x_i}\middle|\mathbf{x}\right] = \frac{\partial F(\mathbf{x})}{\partial x_i} + \frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_j \partial x_j \partial x_i}. \tag{18}$$

Using similar technique as in multiplicative noise, we can get that

$$E[F(\tilde{\mathbf{x}})|\mathbf{x}] = F(\mathbf{x}) + \frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} \tag{19}$$

and

$$\frac{\partial}{\partial x_i}E[F(\tilde{\mathbf{x}})|\mathbf{x}] = \frac{\partial F(\mathbf{x})}{\partial x_i} + \frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^3 F(\mathbf{x})}{\partial x_i \partial x_j \partial x_j}. \tag{20}$$

Compare (18) with (20), we get that $E\left[\partial F(\tilde{\mathbf{x}})/\partial x_i \middle| \mathbf{x}\right] = \partial E[F(\tilde{\mathbf{x}})|\mathbf{x}]/\partial x_i$ and thus

$$\frac{\partial V_{\otimes}(\mathbf{x}(t))}{\partial \mathbf{x}} = \frac{\partial E[F(\tilde{\mathbf{x}})|\mathbf{x}]}{\partial \mathbf{x}} + \lambda \mathbf{x} \tag{21}$$

By (19), (20) and (21), the energy function as stated in (17) can be obtained and the proof is completed. **Q.E.D.**

3.2 Chaotic Noise

For the system with chaotic noise injection, all elements in $\mathbf{x}$ suffered the same amount of noise $\kappa n(t)$ in the t^{th} step. As observed from Fig. 1 which plots the value $\sum_{\tau=0}^{T-1} n(t+\tau)/T$ for $t = 1, \cdots, 2000$ and for different values of T, it is reasonable to assume that $\sum_{\tau=0}^{T-1} n(t+\tau)/T$ is a constant for all t if $T \gg 1$. Then, we can get the following theorem on the energy function of a gradient system with chaotic noise.

Theorem 3. *For a gradient system defined as (5) and* $\mu T \to 0$,

$$V_{\odot}(\mathbf{x}) = F(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \kappa' \sum_{i=1}^{n} x_i, \tag{22}$$

where κ' *is a constant.*

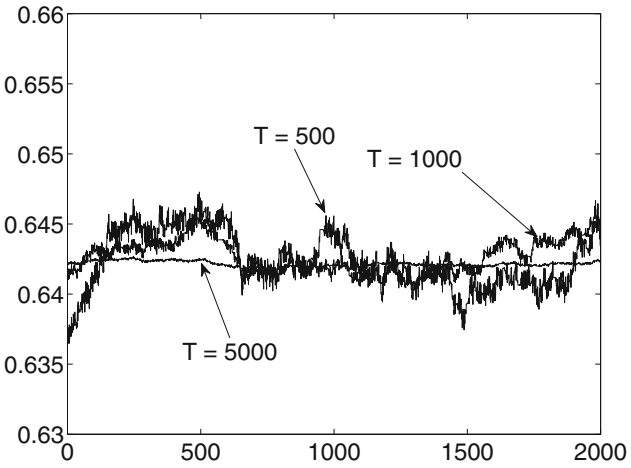

Fig. 1. $\sum_{\tau=0}^{T-1} n(t+\tau)/T$ against t, for $t = 1, \cdots, 2\,K$; $T = 500, 1\,K, 5\,K$. $\alpha = 3.8$

Proof: Suppose $\mu T \to 0$ for all t, we could assume that

$$\mathbf{x}(t+\tau) = \mathbf{x}(t), \quad \frac{\partial F(\mathbf{x}(t+\tau))}{\partial \mathbf{x}} = \frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}}$$

for $\tau = 0, 1, \cdots, T-1$. In such case, we can get from (5) that

$$\begin{aligned} \mathbf{x}(t+T) &= \mathbf{x}(t) - \mu' \left(\frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}} + \lambda \mathbf{x}(t) + \frac{\kappa}{T} \sum_{\tau=0}^{T-1} n(t+\tau)\mathbf{e} \right) \\ &= \mathbf{x}(t) - \mu' \left(\frac{\partial F(\mathbf{x}(t))}{\partial \mathbf{x}} + \lambda \mathbf{x}(t) + \kappa \bar{n} \mathbf{e} \right), \end{aligned} \tag{23}$$

where $\mu' = \mu T$ and $\bar{n} = \lim_{T\to\infty} \sum_{t=1}^{T} n(t)/T$. Clearly, the energy function is given by (22) and $\kappa' = \kappa \lim_{T\to\infty} \sum_{t=1}^{T} n(t)/T$. **Q.E.D.**

4 Effect of Noise

For multiplicative noise, let us rewrite that $V_{\otimes}(\mathbf{x}) = F(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + S_b \mathcal{R}(\mathbf{x})$, where $\mathcal{R}(\mathbf{x})$ corresponds to a regularizer. From (11), it is given by

$$\mathcal{R}(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^{n} \frac{\partial^2 F(\mathbf{x})}{\partial x_j \partial x_j} x_j^2 - \int \mathbf{x} \otimes \mathbf{diag}\left\{\mathbf{H}(\mathbf{x})\right\} \cdot d\mathbf{x}. \tag{24}$$

The effect of the first term is to bring $\mathbf{x}$ closer to the zero vector while the second term is to push it away. Therefore, the existence of multiplicative noise

in a gradient system would lead to two opposite effects. It should also be noted that $\mathbf{H}(\mathbf{x})$ is a constant matrix (say $\bar{\mathbf{H}}$) and $\mathcal{R}(\mathbf{x}) = 0$ if $F(\mathbf{x})$ is quadratic. Existence of multiplicative noise has no effect on the gradient system.

For additive noise, the additional term $(S_b/2)\sum_{j=1}^{n} \partial^2 F(\mathbf{x})/\partial x_j \partial x_j$ has the effect that brings the solution closer to the zeros vector. This term reduces to a constant if $F(\mathbf{x})$ is quadratic. The different between $V_{\oplus}(\mathbf{x})$ and $V(\mathbf{x})$ is just a constant. So, existence of additive noise has no effect on a gradient system if $F(\mathbf{x})$ is quadratic.

For chaotic noise, from (22), the additional term is $\kappa' \sum_{i=1}^{n} x_i$. Its effect is to let $\mathbf{x}$ slide along the direction $-[1\ 1\ \cdots\ 1]^T$. If all the x_is are positive, the additional term will bring them move slightly towards the zero vector. If all the x_i are negative, it will move $\mathbf{x}$ slightly further away from the zero vector. The effect of noise on the gradient system will depend on the minimum point of the $V(\mathbf{x})$ and it has no effect if $\lim_{T\to\infty} \sum_{t=1}^{T} n(t)/T = 0$.

5 Conclusion

In this paper, we have introduced the models of the gradient systems with three different type of noise, namely multiplicative, additive and chaotic noise. The energy functions of the corresponding gradient systems with noise have been revealed. By investigating the additional term in the energy functions as compared with the original energy function, it is found that only additive noise has a clear effect on the gradient system. It enforces the state vector moving slightly towards the zero vector. With multiplicative noise, two opposite effects exists, moving towards and away. With chaotic noise, the effect will be depended on the location of the minimum point of $V(\mathbf{x})$. It could be enforced to move towards or away from the zero vector. Moreover, if $F(\mathbf{x})$ is quadratic, either multiplicative or additive noise will have no effect on the gradient system.

Treating (i) the state vector as the weight vector of a neural network, (ii) $F(\cdot)$ as the mean square errors and (iii) moving toward the zero vector as improving generalization, our results imply that (a) injecting AWN during MLP learning can improve generalization, (b) injecting MWN or CN during MLP learning might not be and (c) injecting AWN or MWN during RBF learning cannot improve generalization. Results (a) and (b) are equally applied to other nonlinear neural networks. Treating x as the state variable in the stochastic Wang's kWTA model [18] or $\mathbf{x}$ as the neuronal outputs of Hopfield network, the effect of noise on these models can thus be analyzed by the same technique. Due to page limit, those results will be presented elsewhere.

Acknowledgments. 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. An, G.: The effects of adding noise during backpropagation training on a generalization performance. Neural Comput. **8**, 643–674 (1996)
2. Asai, H., Onodera, K., Kamio, T., Ninomiya, H.: A study of Hopfield neural networks with external noises. In: Proceedings IEEE International Conference on Neural Networks, vol. 4, pp. 1584–1589 (1995)
3. Azamimi, A., Uwate, Y., Nishio, Y.: An analysis of chaotic noise injected to backpropagation algorithm in feedforward neural network. In: Proceedings of IWVCC08, pp. 70–73 (2008)
4. Azamimi, A., Uwate, Y., Nishio, Y.: Effect of chaos noise on the learning ability of back propagation algorithm in feed forward neural network. In: Proceedings of the 6th Internationa Colloquium on Signal Processing and Its Applications (CSPA) (2010)
5. Bishop, C.M.: Training with noise is equivalent to Tikhonov regularization. Neural Comput. **7**, 108–116 (1995)
6. Bolt, G.: Fault tolerant in multi-layer Perceptrons. Ph.D. Thesis, University of York, UK (1992)
7. Grandvalet, Y., Canu, S.: A comment on noise injection into inputs in backpropagation learning. IEEE Trans. Syst. Man Cybern. **25**, 678–681 (1995)
8. Grandvalet, Y., Canu, S., Boucheron, S.: Noise injection: theoretical prospects. Neural Comput. **9**, 1093–1108 (1997)
9. Ho, K., Leung, C.S., Sum, J.: Convergence and objective functions of some fault/noise injection-based online learning algorithms for RBF networks. IEEE Trans. Neural Netw. **21**(6), 938–947 (2010)
10. Ho, K., Leung, C.S., Sum, J.: Objective functions of the online weight noise injection training algorithms for MLP. IEEE Trans. Neural Netw. **22**(2), 317–323 (2011)
11. Jim, K.C., Giles, C.L., Horne, B.G.: An analysis of noise in recurrent neural networks: convergence and generalization. IEEE Trans. Neural Netw. **7**, 1424–1438 (1996)
12. Murray, A.F., Edwards, P.J.: Synaptic weight noise during multilayer perceptron training: fault tolerance and training improvements. IEEE Trans. Neural Netw. **4**(4), 722–725 (1993)
13. Murray, A.F., Edwards, P.J.: Enhanced MLP performance and fault tolerance resulting from synaptic weight noise during training. IEEE Trans. Neural Netw. **5**(5), 792–802 (1994)
14. Reed, R., Marks II, R.J., Oh, S.: Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter. IEEE Trans. Neural Netw. **6**(3), 529–538 (1995)
15. Sequin, C.H., Clay, R.D.: Fault tolerance in feedforward artificial neural networks. Neural Netw. **4**, 111–141 (1991)
16. Sum, J., Leung, C.S., Ho, K.: Convergence analysis of on-line node fault injection-based training algorithms for MLP networks. IEEE Trans. Neural Netw. Learn. Syst. **23**(2), 211–222 (2012)
17. Sum, J., Leung, C.S., Ho, K.: Convergence analyses on on-line weight noise injection-based training algorithms for MLPs. IEEE Trans. Neural Netw. Learn. Syst. **23**(11), 1827–1840 (2012)
18. Sum, John, Leung, Chi-sing, Ho, Kevin: Effect of input noise and output node stochastic on Wang's kWTA. IEEE Trans. Neural Netw. Learn. Syst. **24**(9), 1472–1478 (2013)

19. Wang, L.: Noise injection into inputs in sparsely connected Hopfield and winner-take-all neural networks. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **27**(5), 868–870 (1997)
20. Zhang, H., Zhang, Y., Xu, D., Liu, X.: Deterministic convergence of chaos injection-based gradient method for training feedforward neural networks, to appear in Cognitive Neurodynamic