Effect of Noise on Gradient Systems

Kevin Ho, Hsia-Ching Chang, and Wei-Bin Lee

***Abstract*—Recently, we have analyzed the convergence properties and derived the objective functions of the weight noise injection-based algorithms. In this paper, we generalize our work to perturbed gradient systems (PGS). For original gradient systems (GS), the change of a vector *x*(*t*) at time t is defined as the gradient vector of a function *F*(*x*) times a small negative constant. For PGS, it is assumed that the vector *x*(*t*) is perturbed by mean zero Gaussian noise. The noise can either be multiplicative or additive. In this paper, the corresponding energy functions for both cases are derived. It is found that their energy functions are very difference from *F*(*x*). For the case of multiplicative noise, the energy function consists of three terms: (i) *F*(*x*), (ii) a regularization term and (iii) a de-regularization term. For the case of additive noise, the energy function consists of only two terms: (i) *F*(*x*) and (ii) a regularization term. Note that de-regularization could lead to divergence behavior while regularization will improve the convergence behavior of a system. Our results suggest that special caution should be done to a gradient system with multiplicative noise.**

***Index Terms*—Additive noise, energy function, gradient systems, multiplicative noise, noise effect.**

I. Introduction

In this paper, we focus on a more general topic which is on the effect of noise on gradient systems. It is because many learning algorithms (like back propagation and PCA learning) and neural network models (like associative memory and *k*WTA) can be modeled as gradient systems. Their dynamical behaviors are equivalent to energy minimization.

Understanding the effect of noise on gradient systems can aid to the understanding of the effect of noise on learning algorithms and neural dynamics.

Research on the effect of noise on neural networks has been conducted for almost two decades. In particular in the 1990s, some researchers were actively in investigating the effect of noise on the learning algorithms for multilayer perceptrons [1]-[4], the effect of noise on recurrent neural network [5], and the effect of noise on the dynamical behaviors of associative networks [6]. While the aforementioned researches relied on simulation studies, some other researchers conducted theoretical analysis. But their research results are limited to the effect of *additive input noise* on the learning algorithms for multilayer perceptrons [7]-[11]. Until recent years, the analysis on the effect of *additive and multiplicative weight noise* on the learning algorithms for multilayered perceptrons has been done [12]-[15]. Moreover, the analysis on the effect of noise on *k*WTA has been proceeded [16].

Since many neural network learning algorithms and the dynamical behavior of many recurrent networks can be modeled as gradient systems, analysis on the effect of noise in these learning algorithms and the dynamic behaviors from gradient systems perspective can provide a more general picture (and even a unified theory) for further investigation on other complex neural models and learning algorithms which have not yet studied.

II. Gradient Systems with Noise

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

$$x(t+1) = x(t) - \mu(t)\frac{\partial F(x(t))}{\partial x} \tag{1}$$

where $\mu(t) > 0$, $\mu(t) \to 0$ is the step size at the t^{th} step and

$$\frac{\partial F(x(t))}{\partial x} = \frac{\partial F(x)}{\partial x}\bigg|\ x{=}x(t)$$

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

$$\tilde{x}(t) = x(t) + b(t) \otimes x(t) \tag{2}$$

In (2), $b(t) \in R^n$ is a Gaussian random vector with mean 0 and covariance matrix $S_b I_{n\times n}$. $\otimes$ is a element-wise multiplication operator, i.e.

$$b(t) \otimes x(t) = (b_1(t)x_1(t), b_2(t)x_2(t), ..., b_n(t)x_n(t))^T$$

The gradient system (1) is given as follows:

$$x(t+1) = \tilde{x}(t) - \mu(t)\frac{\partial F(\tilde{x}(t))}{\partial x} \tag{3}$$

Here, we assume that $E[b_i(t)] = 0$ for all $i = 1,\dots, n$ and $t \ge 0$. $E[b_i(t)b_j(t)]$ equals zero if $i \neq j$; otherwise, S_b. And, $E[b_i(t_1)b_j(t_2)]$=0 if $t_1 \neq t_2$.

III. Main Results

Given $x(t)$, we get the mean update of (3) that

$$E\Big[x(t+1)\big|x(t)\Big] = E\Big[\tilde{x}(t)\big|x(t)\Big] - \mu(t)E\left[\frac{\partial F(\tilde{x}(t))}{\partial x}\big|x(t)\right] \tag{4}$$

Manuscript received November 17, 2012; revised April 2, 2013.

Kevin Ho is with the Department of Computer Science and Communication Engineering, Providence University, Sha-Lu, Taiwan (e-mail: ho@pu.edu.tw).

Hsia-Ching Chang is with the Department of Information Management, Hsiuping University of Science and Technology, Dali Dist., Taichung, Taiwan ROC (e-mail: hcchang@mail.hust.edu.tw).

Wei-Bin Lee is with the Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan ROC (e-mail: wblee@fcu.edu.tw).

DOI: 10.7763/IJCTE.2013.V5.808

In (4), the expectation is taken over the probability space of $\tilde{x}(t)$. Since $E[b(t)]=\mathbf{0}$, by (2) we get that $E\left[\tilde{x}(t)\middle|x(t)\right] = x(t)$. Equation (4) can be rewritten as follows:

$$E\left[x(t+1)\middle|x(t)\right]=E\left[\tilde{x}(t)\middle|x(t)\right]-\mu(t)E\left[\frac{\partial F(\tilde{x}(t))}{\partial x}\middle|x(t)\right] \quad (5)$$

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

$$E\left[x(t+1)\middle|x(t)\right]=x(t)-\mu(t)\frac{\partial V_{\otimes}(x(t))}{\partial x} \quad (6)$$

Then we can have the following theorem.

Theorem 1: For a gradient system defined as (1) and $x(t)$ is corrupted by multiplicative noise as stated in (2),

$$E\left[F(\tilde{x})\middle|x\right]=F(x)+\frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_j}x_j^2 \quad (7)$$

and

$$V_{\oplus}(x)=F(x)+\frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_j}x_j^2$$
$$-S_b\int x\otimes\mathbf{diag}\{H(x)\}\cdot dx \quad (8)$$

where $\int$is the line integral, $H(x)$ is the Hessian matrix of $F(x)$, i.e. $H(x)=\nabla\nabla_x F(x)$ and

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

Proof: Consider (5) and let $\frac{\partial F(x)}{\partial x_i}$ be the i^{th} element of $\frac{\partial F(x)}{\partial x}$

$$\frac{\partial F(\tilde{x})}{\partial x_i}=\frac{\partial F(x)}{\partial x_i}+\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_i}(b_j x_j)$$
$$+\frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\partial^3 F(x)}{\partial x_k\partial x_j\partial x_i}b_k b_j x_k x_j \quad (9)$$

therefore

$$E\left[\frac{F(\tilde{x})}{\partial x_i}\middle|x\right]=\frac{\partial F(x)}{\partial x_i}+\frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^3 F(x)}{\partial x_j\partial x_j\partial x_i}x_j^2 \quad (10)$$

On the other hand,

$$F(\tilde{x})=F(x)+\sum_{i=1}^{n}\frac{\partial F(x)}{\partial x_i}b_i x_i$$
$$+\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_i}b_j b_i x_j x_i \quad (11)$$

thus

$$E\left[F(\tilde{x})\middle|x\right]=F(x)+\frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_j}x_j^2 \quad (12)$$

and

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

Based on the fact that

$$\frac{\partial^3 F(x)}{\partial x_j\partial x_j\partial x_i}=\frac{\partial^3 F(x)}{\partial x_i\partial x_j\partial x_j} \quad (14)$$

and $F(x)$ is differentiable up to the third degree. Comparing (10) and (13), we get that

$$E\left[\frac{F(\tilde{x})}{\partial x_i}\middle|x\right]=\frac{\partial}{\partial x_i}E\left[F(\tilde{x})\middle|x\right]-S_b\frac{\partial^2 F(x)}{\partial x_i\partial x_i}x_i \quad (15)$$

Further, by (5) and (6), we get that

$$V_{\oplus}(x)=E\left[F(\tilde{x})\middle|x\right]-S_b\int x\otimes\mathbf{diag}\{H(x)\}\cdot dx \quad (16)$$

In other words,

$$V_{\oplus}(x)=F(x)+\frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_j}x_j^2$$
$$-S_b\int x\otimes\mathbf{diag}\{H(x)\}\cdot dx \quad (17)$$

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

Let us write that $V_{\oplus}(x)=F(x)+S_b R(x)$, where $R(x)$ corresponds to a regularizer. From (8), it is given by

$$R(x)=\frac{1}{2}\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_j}x_j^2-\int x\otimes\mathbf{diag}\{H(x)\}\cdot dx \quad (18)$$

The effect of the first term is to bring x closer to the zero vector while the second term is to push it away from the zero vector. Therefore, the existence of multiplicative noise in a gradient system would lead to both regularization effect and de-regularization effect. This effect does not exist if the noise is additive (see below).

It should be noted that $H(x)$ is a constant matrix (say $\bar{H}$) if $F(x)$ is quadratic. One can easily show that $R(x)=0$. Thus, we can state without proof the following corollary.

Corollary 1*:* For a gradient system defined as (1) in which $x(t)$ is corrupted by multiplicative noise as stated in (2) and $F(x)$ is quadratic, $V_{\otimes}(x)=F(x)$.

For the system which is corrupted by additive noise,

$$\tilde{x}(t)=x(t)+b(t) \quad (19)$$

where $b(t)\in R^n$ is a Gaussian random vector with mean 0 and covariance matrix $S_b I_{n\times n}$. Then, we can have the following theorem.

Theorem 2: For a gradient system defined as (1) and $x(t)$ is corrupted by additive noise

$$V_{\oplus}(x)=F(x)+\frac{S_b}{2}\sum_{j=1}^{n}\frac{\partial^2 F(x)}{\partial x_j\partial x_j} \quad (20)$$

Proof: For additive noise, the noisy x in (3) is given by $\tilde{x} = x + b$. Similarly, we consider (5) and let $\frac{\partial F(x)}{\partial x_i}$ be the i^{th} element of $\frac{\partial F(x)}{\partial x}$.

$$\frac{\partial F(\tilde{x})}{\partial x_i} = \frac{\partial F(x)}{\partial x_i} + \sum_{j=1}^{n} \frac{\partial^2 F(x)}{\partial x_j \partial x_i} b_j + \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \frac{\partial^3 F(x)}{\partial x_k \partial x_j \partial x_i} b_k b_j \qquad (21)$$

therefore

$$E\left[\frac{F(\tilde{x})}{\partial x_i} \middle| x \right] = \frac{\partial F(x)}{\partial x_i} + \frac{S_b}{2} \sum_{j=1}^{n} \frac{\partial^3 F(x)}{\partial x_j \partial x_j \partial x_i} \qquad (22)$$

On the other hand,

$$F(\tilde{x}) = F(x) + \sum_{i=1}^{n} \frac{\partial F(x)}{\partial x_i} b_i + \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 F(x)}{\partial x_j \partial x_i} b_j b_i \qquad (23)$$

thus

$$E\left[F(\tilde{x}) \middle| x \right] = F(x) + \frac{S_b}{2} \sum_{j=1}^{n} \frac{\partial^2 F(x)}{\partial x_j \partial x_j} \qquad (24)$$

and

$$\frac{\partial}{\partial x_i} E\left[F(\tilde{x}) \middle| x \right] = \frac{\partial F(x)}{\partial x_i} + \frac{S_b}{2} \sum_{j=1}^{n} \frac{\partial^3 F(x)}{\partial x_i \partial x_j \partial x_j} \qquad (25)$$

By (14) and comparing (22) with (25), we get that

$$E\left[\frac{F(\tilde{x})}{\partial x_i} \middle| x \right] = \frac{\partial}{\partial x_i} E\left[F(\tilde{x}) | x \right] \qquad (26)$$

As a result,

$$V_{\oplus}(x) = E\left[F(\tilde{x}) | x \right] = F(x) + \frac{S_b}{2} \sum_{j=1}^{n} \frac{\partial^2 F(x)}{\partial x_j \partial x_j} \qquad (27)$$

The additional term has the effect that brings the solution closer to the zeros vector. The proof is completed Q.E.D.

IV. Illustrative Examples

In this section, we illustrate by two examples the application of Theorem 1.

A. *Simple Gradient System with Multiplicative Noise*

Consider a simple example that $x \in R$. Let a gradient system is with objective function $F(x)$ given by

$$F(x) = x^4 - 3x^3 - 3x^2 + 5x \qquad (28)$$

Its shape is shown in Fig. 1. The noise-free update can then be expressed as follows

$$x(t+1) = x(t) - \mu(t) F^{'}(x(t)) \qquad (29)$$

where $F^{'}(x) = 4x^3 - 9x^2 - 6x + 5$. The update of x(t) is given by

$$x(t+1) = \tilde{x}(t) - \mu(t)(4\tilde{x}^3(t) - 9\tilde{x}^2(t) - 6\tilde{x}(t) + 5) \qquad (30)$$

where $\tilde{x}(t) = x(t) + b(t)x(t)$ and $b(t)$ is a mean zero Gaussian noise with variance S_b. For S_b is small, we can get from (7) that

$$E\left[F(\tilde{x}) | x \right] = F(x) + S_b(6x^4 - 9x^3 - 3x^2) \qquad (31)$$

The de-regularization term in (8) will be given by

$$S_b \int x \frac{d^2 F(x)}{dx^2} dx = S_b(3x^4 - 6x^3 - 3x^2) \qquad (32)$$

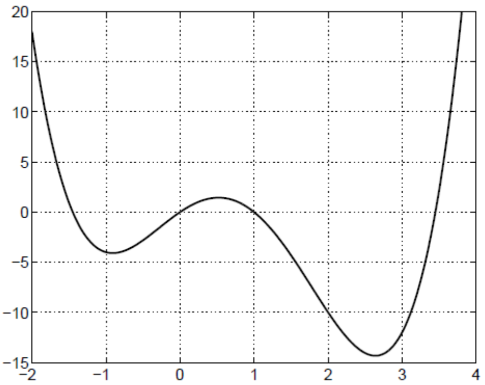


Fig. 1. The shape of $F(x)$.

Thus, the objective function of (30) is that

$$V(x) = F(x) + 3S_b x^2 (x - 3) \qquad (33)$$

Fig. 2 shows the locations of the minimums of the functions $F(x)$, $E\left[F(\tilde{x}) | x \right]$ and $V(x)$ for S_b=0.01. It is clear that the minimum points of $V(x)$ lie in between the minimum points of $F(x)$ and $E\left[F(\tilde{x}) | x \right]$.

B. *Stochastic Wang's* kWTA *with Multiplicative Noise*

Stochastic Wang's *k*WTA is defined as follows [13]:

$$x(t+\tau) = x(t) + \mu(t) \left\{ \sum_{i=1}^{n} f(u_i - x(t)) - k \right\} \qquad (34)$$

where $u_1, \ldots, u_n$ are the inputs to the network, k is a positive integer and β is the step size. In (34),

$$f(u_i - x(t)) = \frac{1}{1 + \exp(-\alpha(u_i - x(t))} \qquad (35)$$

which is the firing rate of the i^{th} neuron. Moreover, we have shown that (34) is a gradient system,

$$x(t+\tau) = x(t) + \mu(t) \frac{dF(x(t))}{dx} \qquad (36)$$

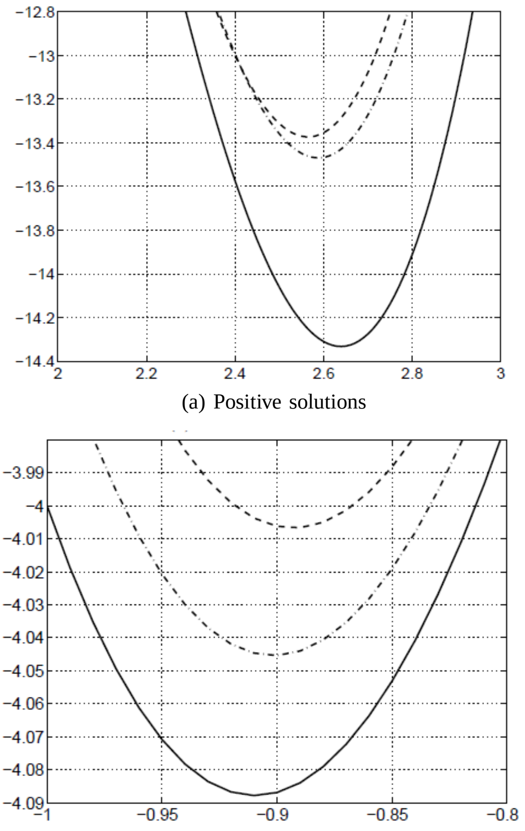
with energy function

$$F(x) = kx + \alpha^{-1} \sum_{i=1}^{n} \log(1 + \exp(\alpha(u_i - x))) \qquad (37)$$

Now, suppose the state variable $x(t)$ is corrupted by multiplicative noise, i.e. $\tilde{x}(t) = x(t) + b(t)x(t)$. The state-space model for this stochastic Wang's *k*WTA is given by

$$x(t+\tau) = x(t) + \mu(t)\left\{\sum_{i=1}^{n} f(u_i - \tilde{x}(t)) - k\right\} \qquad (38)$$

By Theorem 1, the energy function of (38) can be obtained.



(a) Positive solutions

(b) Negative solutions

Fig. 2. Solution points of the functions $F(x)$ (solid), $E\left[F(\tilde{x})\middle|x\right]$ (dash) and $V(x)$ (dot-dash).

From (34) and (36), we can get that

$$\frac{d^2F(x)}{dx^2} = -n\sum_{i=1}^{n}\frac{df(u_i - x)}{dx}$$

$$= \alpha\sum_{i=1}^{n} f(u_i - x)(1 - f(u_i - x)) \qquad (39)$$

Besides

$$\int x\frac{d^2F(x)}{dx^2}dx = x\frac{dF(x)}{dx} - F(x)$$

Therefore, the energy function of (38) is given as follows:

$$V(x) = (1 + S_b)F(x)$$

$$+\frac{\alpha S_b}{2}x^2\sum_{i=1}^{n} f(u_i - x)(1 - f(u_i - x))$$

$$-S_b x\left(k - \sum_{i=1}^{n} f(u_i - x)\right) \qquad (40)$$

where $F(x)$ in (40) is the same as the one given by (37).

V. Conclusion

In this paper, we have introduced gradient systems which are corrupted by either additive or multiplicative noise. Given that the energy function of the original gradient system is $F(\mathrm{x})$, we have shown that the energy function of the noise corrupted gradient system (denoted by $V(\mathrm{x})$) is given by $V(\mathrm{x}) = F(\mathrm{x}) + S_b\,\boldsymbol{R}(\mathrm{x})$, where $\boldsymbol{R}(\mathrm{x})$ is the regularizer as stated in (18). Applications of the analytical results are illustrated by a simple gradient descent system and the stochastic Wang's *k*WTA.

Acknowledgment

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] G. Bolt, *Fault Tolerant in Multi-Layer Perceptrons*, PhD Thesis, University of York, UK, 1992.

[2] 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, no. 4, pp. 722-725, 1993.

[3] 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, no. 5, pp. 792-802, 1994.

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

[5] 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, pp. 1424-1438, 1996.

[6] 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, no. 5, pp. 868-870, October, 1997.

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

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

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

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

[11] R. Reed, R. J. Marks, and S. Oh, "Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter," *IEEE Transactions on Neural Networks*, vol. 6, no. 3, pp. 529-538, 1995.

[12] 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, no. 6, pp. 938-947, June, 2010.

[13] K. Ho, C. S. Leung, and J. Sum, "Objective functions of the online weight noise injection training algorithms for MLP," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 317-323, Feb 2011.

[14] J. Sum, C. S. Leung, and K. Ho, "Convergence analysis of on-line node fault injection-based training algorithms for MLP networks," *IEEE Transac- tions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 211-222, Feb. 2012.

[15] J. Sum, C. S. Leung, and 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, no. 11, pp. 1827-1840, Nov. 2012.

[16] J. Sum, C. S. Leung, and K. Ho, "Analysis on Wang's *k*WTA with stochastic behaviors," in *Proceedings of ICONIP 2011*, Shanghai, November 13-17, 2011.

Kevin I-J Ho received his B.S. in computer engineering from the National Chiao-Tung University, Taiwan in 1983. From 1985 to 1987, he was an assistant engineer of the Institute of Information Industry, Taiwan. Then, he received his M.S. and Ph.D. in computer science from the University of Texas at Dallas in 1990 and 1992 respectively. Currently, he is a professor of the Department of Computer Science and Communication Engineering, Providence University, Taiwan. His current research interests include neural computation, algorithm design and analysis, security and scheduling theory.

Hsia-Ching Chang is currently a lecturer of the Department of Information Management, Hsiuping University of Science and Technology. She received her master in management science from Providence University, Taiwan.

Wei-Bin Lee received his B.S degree from the Department of Information and Computer Engineering, Chung-Yuan Christian University, Chungli, Taiwan, in 1991 and his M.S. degree in computer science and information engineering from the National Chung Cheng University, Chiayi, Taiwan in 1993. He received his Ph.D. degree in 1997 from the National Chung Cheng University. Since 1999, he has been with the Department of Information Engineering and Computer Science at the Feng Chia University, where he is currently a professor. He also serves as the Dean of the Office of Information Technology, and the Director of Information and Communication Security Research Center at the Feng Chia University. His research interests currently include cryptography, information security, digital rights management, steganography, medical information security, and security of electronic medical record. He is an honorary member of the Phi Tau Phi Scholastic Honor Society.