Proceedings of the 2007 IEEE Symposium on
Foundations of Computational Intelligence (FOCI 2007)

Convergence of Online Gradient Algorithm with Stochastic Inputs for Pi-Sigma Neural Networks*

Xidai Kang, Yan Xiong, Chao Zhang, Wei Wu

Department of Applied Mathematics, Dalian University of Technology,
Dalian 116024, P.R. China
wuweiw@dlut.edu.cn

Abstract. An online gradient method is presented and discussed for Pi-Sigma neural networks with stochastic inputs. The error function is proved to be monotone in the training process, and the gradient of the error function tends to zero if the weights sequence is uniformly bounded. Furthermore, after adding a moderate condition, the weights sequence itself is also proved to be convergent.

Key words: Pi-Sigma neural network, online gradient algorithm, stochastic input, convergence, monotonicity.

1 Introduction

Pi-Sigma Network (PSN) [4] is a class of higher order feedforward polynomial neural network and is known to provide inherently more powerful mapping abilities than traditional feedforward neural networks. The neural networks consisting of the PSN modules are widely used for classification and approximation problems [2, 5]. The online gradient method is usually used to trained PSN. A convergence analysis of the online gradient method (OGM for short) with fixed order inputs (OGM-F) for PSN is presented in [6]. The aim of this paper is to generalize the result in [6] to the online gradient method with special stochastic inputs (OGM-SS) for PSN. The motivation for us to make such an extension is the following. Apart from the computational efficiency, another reason for people to choose OGM rather than the ordinary gradient method is that OGM helps for the iteration procedure to jump off from local minima due to its stochastic nature (see for instance ([1])). But in OGM-F, the stochastic nature is somehow lost, since its iteration procedure is completely determined as long as the order of the input samples is fixed. OGM-SS recovers the stochastic nature and hence is an important improvement of OGM-F.

* The research is partly supported by the National Natural Science Foundation of China (10471017)

2 PSN and OGM-SS

Let the numbers of neurons for the input, summation and product layers of the PSN be P, N and 1, respectively. We denote by $w_k = (w_{k1}, w_{k2}, \cdots, w_{kP})^T$ $(1 \leq k \leq N)$ the weight vector connecting the summation node k and the input nodes, and write $w = (w_1^T, w_2^T, \cdots, w_N^T)^T \in R^{NP}$. The weights on the connections between the product node and the summation nodes are fixed to 1. We have included a special input unit ξ_P, corresponding to the biases w_{kP}, with fixed value -1. The topological structure of PSN is shown in Fig. 1.

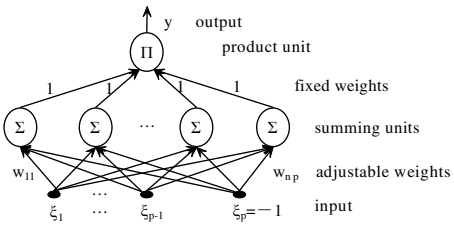

Fig. 1. PSN structure with a single output.

Supply the network with a training example set $\{\xi^j, O^j\}_{j=1}^J \subset \mathbb{R}^P \times \mathbb{R}$, where $\xi^j = (\xi_1^j, \cdots, \xi_{P-1}^j, \xi_P^j) \in \mathbb{R}^P$ and $\xi_P^j \equiv -1$. Assume $g : \mathbb{R} \to \mathbb{R}$ is a given activation function. For an input $x \in \mathbb{R}^P$, the output of the network is

$$y = g(\prod_{i=1}^{N}(w_i \cdot x)) \tag{1}$$

For simplicity, we write $g_j(t) = \frac{1}{2}(O^j - g(t))^2$. Define the error function as

$$E(w) = \frac{1}{2}\sum_{j=1}^{J}(O^j - y^j)^2 = \frac{1}{2}\sum_{j=1}^{J}\Big(O^j - g(\prod_{i=1}^{N}(w_i \cdot \xi^j))\Big)^2 = \sum_{j=1}^{J} g_j(\prod_{i=1}^{N}(w_i \cdot \xi^j)) \tag{2}$$

Its gradient with respect to the weight vector w_k $(k = 1, 2, \cdots, N)$ is

$$E_k(w) = \sum_{j=1}^{J} g_j'(\prod_{i=1}^{N}(w_i \cdot \xi^j))(\prod_{\substack{i=1 \\ i \neq k}}^{N}(w_i \cdot \xi^j))\xi^j \tag{3}$$

In each epoch m of iteration, the set of samples is re-arranged in a stochastic order $m_1, m_2, \cdots, m_J$, which is a permutation of the index set $1, 2, \cdots, J$. For any initial w^0, OGM-SS for PSN modifies the weights in the following manner:

$$w_k^{mJ+j} = w_k^{mJ+j-1} + \triangle_{m_j} w_k^{mJ+j-1},\ 1 \leq j \leq J,\ 1 \leq k \leq N,\ m = 0, 1, \cdots \tag{4}$$

$$\triangle_l w_k^m = -\eta_m g_l'\big(\prod_{i=1}^{N}(w_i^m \cdot \xi^l)\big)\big(\prod_{\substack{i=1\\ i\neq k}}^{N}(w_i^m \cdot \xi^l)\big)\xi^l \tag{5}$$

Here, η_m is a learning rate and β is a positive constant, satisfying

$$1/\eta_{m+1} = 1/\eta_m + \beta,\ m = 0, 1, \cdots \tag{6}$$

Obviously, there exists a constant $\gamma > 0$ such that $\gamma/m < \eta_m < 1/m\beta$.

3 The Main Theorem

The following conditions will be used in this paper (C_0 is a positive constant):

$(A1)$ $|g_j(t)|, |g_j'(t)|, |g_j''(t)| \leq C_0,\ \forall t \in \mathbb{R},\ 1 \leq j \leq J$

$(A2)$ $\|\xi^j\| \leq C_0,\ |w_k^i \cdot \xi^j| \leq C_0,\ \forall 1 \leq j \leq J,\ 1 \leq k \leq N,\ i = 0, 1, \cdots$

$(A3)$ $\{w^i\}_{i=0}^{\infty}$ are contained in bounded closed region $\mathbb{D} \subset \mathbb{R}^{NP}$, and there are finite points in set $\mathbb{D}_0 = \{w \in \mathbb{D} | E_w\{w\} = 0\}$.

Theorem 1. *If* $(A1)$ *and* $(A2)$ *is valid and* $\{w^i\}$ *is generated by* (4)*, then*

$$E(w^{(m+1)J}) \leq E(w^{mJ}),\ m = 0, 1, \cdots \tag{7}$$

$$\lim_{i\to\infty} \|E_k(w^i)\| = 0,\ 1 \leq k \leq N \tag{8}$$

Additionally, if $(A3)$ *is also valid, there exists a* $w^* \in \mathbb{D}_0$ *such that*

$$\lim_{i\to\infty} w^i = w^* \tag{9}$$

4 Proof of Theorem 1

First, we present a few lemmas as preparation to prove Theorem 1. Lemmas 2 amd 3 is can be found in [3]. And the proofs of Lemmas 4 and 5 are similar, though not identical, to the corresponding results in [6]. So the proofs to Lemmas 2-5 below are omitted to save the space.

Lemma 1. *Suppose that the sequence* $\sum_{n=1}^{\infty} \frac{a_n^2}{n} < \infty$*, that* $a_n > 0$ *for* $n = 1, 2, \cdots$*, and that there exists a constant* $\mu > 0$ *satisfying* $|a_{n+1} - a_n| < \frac{\mu}{n}$*. Then, we have* $\lim_{n\to\infty} a_n = 0$.

Lemma 2. *Suppose that* $h : \mathbb{R}^K \longrightarrow \mathbb{R}$ *is continuous and differentiable on a compact set* $D \subset \mathbb{R}^K$*, and that* $\Omega = \{x \in D | \nabla h(x) = 0\}$ *has only finite number of points. If a sequence* $\{z^i\}_{i=1}^{\infty} \subset D$ *satisfies* $\lim_{i\to\infty} \|z^{i+1} - z^i\| = 0$ *and* $\lim_{i\to\infty} \|\nabla h(z^i)\| = 0$*, there exists a point* $z^* \in \Omega$ *such that* $\lim_{i\to\infty} z^i = z^*$.

First, we define

$$r_k^{j,m} = \Delta_{m_j} w_k^{mJ+j-1} - \Delta_{m_j} w_k^{mJ+j},\ 1 \leq j \leq J,\ 1 \leq k \leq N,\ m = 0, 1, \cdots \tag{10}$$

The next lemmas estimate $r_k^{j,m}$ and the change of error.

Lemma 3. *Suppose that* $(A1)$ *and* $(A2)$*are satisfied. Then, there exist constants* $C_1, C_2, C_3 > 0$*, such that for any* $m = 0, 1, \cdots$,

$$\sum_{t=1}^{j} \|r_k^{t,m}\| \leq C_1 \eta_m \sum_{i=1}^{N} \sum_{t=1}^{j} \|\triangle_{m_t} w_i^{mJ}\|,\ 1 \leq j \leq J,\ 1 \leq k \leq N \tag{11}$$

$$\|w_k^{mJ+j} - w_k^{mJ}\| \leq C_2 \sum_{i=1}^{N} \sum_{t=1}^{j} \|\triangle_{m_t} w_i^{mJ}\|,\ 1 \leq j \leq J,\ 1 \leq k \leq N \tag{12}$$

$$|\prod_{\substack{i=1 \\ i \neq k}}^{N} (w_i^{mJ+j} \cdot \xi^l) - \prod_{\substack{i=1 \\ i \neq k}}^{N} (w_i^{mJ} \cdot \xi^l)| \leq C_3 (\sum_{i=1}^{N} \sum_{t=1}^{j} \|\triangle_{m_t} w_i^{mJ}\| + \sum_{i=1}^{N} \sum_{t=1}^{j} \|r_i^{t,m}\|) \tag{13}$$

Lemma 4. *Suppose that* $(A1)$ *and* $(A2)$ *are valid. Then, there exists a constant* α *such that for any* $m = 0, 1, \cdots$,

$$E(w^{(m+1)J}) \leq E(w^{mJ}) - \frac{1}{\eta_m} \sum_{i=1}^{N} \|\sum_{j=1}^{J} \triangle_{m_j} w_i^{mJ}\|^2 + \alpha \sum_{i=1}^{N} \sum_{j=1}^{J} \|\triangle_{m_j} w_i^{mJ}\|^2 \tag{14}$$

Now, we are ready to prove Theorem 1.

Proof to Theorem 1. By Lemma 4, we have for any integer M

$$E(w^{(M+1)J}) \leq E(w^J) + \sum_{m=1}^{M} \sum_{i=1}^{N} \big(\frac{1}{\eta_m} \|\sum_{j=1}^{J} \triangle_{m_j} w_i^{mJ}\|^2 - \alpha \sum_{j=1}^{J} \|\triangle_{m_j} w_i^{mJ}\|^2\big) \tag{15}$$

Note $E(w^{(M+1)J}) \geq 0$ and set $M \to \infty$. Then, using (5) and (7) we get

$$\begin{aligned} &\sum_{m=1}^{\infty} \sum_{i=1}^{N} (\frac{1}{\eta_m} \|\sum_{j=1}^{J} \triangle_{m_j} w_i^{mJ}\|^2) \leq \sum_{m=1}^{\infty} \sum_{i=1}^{N} (\alpha \sum_{j=1}^{J} \|\triangle_{m_j} w_i^{J}\|^2) + E(w^J) \\ &< \sum_{m=1}^{\infty} (\frac{\alpha J N C^{2(N+1)}}{m^2 \beta^2}) + E(w^J) < C_4 \sum_{m=1}^{\infty} \frac{1}{m^2} + E(w^J) < \infty \end{aligned} \tag{16}$$

where $C_4 = \frac{\alpha C^{2(N+1)} JN}{\beta^2}$. Since $\{m_1, m_2, \cdots, m_J\}$ is a permutation of $\{1, 2, \cdots, J\}$, (16) is also valid for $m_i = i, i = 1, 2, \cdots, J$. Using (5) and (6), we obtain

$$\sum_{m=1}^{\infty} \frac{1}{m} \|E_k(w^{mJ})\|^2 < \frac{1}{\gamma} \sum_{m=1}^{\infty} (\frac{1}{\eta_m} \|\sum_{j=1}^{J} \triangle_j w_i^{mJ}\|^2) < \infty \tag{17}$$

From (3), (5), $(A1)$ and $(A2)$, we conclude that

$$\|\triangle_l w_i^{mJ+j}\| = \eta_m \|g_l'\big(\prod_{i=1}^{N} (w_i^{mJ+j} \cdot \xi^l)\big) \prod_{\substack{i=1 \\ i \neq k}}^{N} (w_i^{mJ+j} \cdot \xi^l) \xi^l\| \leq C^{N+1} \eta_m < \frac{C_5}{m} \tag{18}$$

where $C_5 = C^{n+1}/\beta$. By the mean value theorem we can easily get that

$$\begin{aligned} & g_l'(\prod_{i=1}^{N}(w_i^{mJ+j}\cdot\xi^l)) - g_l'(\prod_{i=1}^{N}(w_i^{mJ}\cdot\xi^l)) \\ & = g''(t_{j,m,l})\sum_{k=1}^{N}\big(\prod_{\substack{i=1\\ i\neq k}}^{N}(\widetilde{t}_{i,m,j,l})(w_k^{mJ+j}-w_k^{mJ})\cdot\xi^l\big) \end{aligned} \tag{19}$$

where $t_{j,m,l}$ and $\widetilde{t}_{i,m,j,l}$ are suitable constants. Let $e\in\mathbb{R}^P$ be an arbitrary unit vector. (3) and (19) lead to

$$\begin{aligned} & E_k(w^{mJ+j})\cdot e - E_k(w^{mJ})\cdot e \\ & = \sum_{l=1}^{J}\Big(g_l'(\prod_{i=1}^{N}(w_i^{mJ}\cdot\xi^l))(\prod_{\substack{i=1\\ i\neq k}}^{N}(w_i^{mJ+j}\cdot\xi^l) - \prod_{\substack{i=1\\ i\neq k}}^{N}(w_i^{mJ}\cdot\xi^l)) + \\ & g_l''(t_{j,m,l})\prod_{\substack{i=1\\ i\neq k}}^{N}(w_i^{mJ+j}\cdot\xi^l)\sum_{k_1=1}^{N}\big((\prod_{\substack{i=1\\ i\neq k_1}}^{N}\widetilde{t}_{i,m,j,l})(w_{k_1}^{mJ+j}-w_{k_1}^{mJ})\cdot\xi^l\big)(\xi^l\cdot e)\Big) \end{aligned} \tag{20}$$

So from $(A1)$, (11), (12), (13), (18) and Cauchy-Schwartz inequality, we have

$$\begin{aligned} & \big||E_k(w^{mJ+j})\cdot e| - |E_k(w^{mJ})\cdot e|\big| \le C^N\sum_{l=1}^{J}\Big(\big(\sum_{i=1}^{N}\sum_{t=1}^{j}\|\triangle_{m_t}w_i^{mJ}\| + \\ & \sum_{i=1}^{N}\sum_{t=1}^{j}\|r_i^{t,m}\|\big)|\xi^l\cdot e|\Big) + C^{2N-1}\sum_{l=1}^{J}\big(|\xi^l\cdot e|\sum_{k_1=1}^{N}|(w_{k_1}^{mJ+j}-w_{k_1}^{mJ})\cdot\xi^l|\big) \\ & < \big(C^{N+1}J(1+C_3N\eta_0)+C^{2N+1}JNC_4\big)NJ(\frac{C_5}{m}) = \frac{C_6}{m} \end{aligned} \tag{21}$$

where $C_6 = (C^{N+1}J(1+C_3N\eta_0)+C^{2N+1}JNC_4)NJC_5$. Using (17) and Cauchy-Schwartz inequality, we obtain that

$$\sum_{m=1}^{\infty}\frac{1}{m}|E_k(w^{mJ})\cdot e|^2 \le \sum_{m=1}^{\infty}\frac{1}{m}\|E_k(w^{mJ})\|^2 < \infty,\ 1\le k\le N \tag{22}$$

Then by (21), (22) and Lemma 1, we conclude that

$$\lim_{m\to\infty}|E_k(w^{mJ})\cdot e| = 0,\ 1\le k\le N \tag{23}$$

It follows from (21) that for $1\le k\le N,\ 1\le j\le J,\ m=1,2,\cdots$

$$|E_k(w^{mJ+j})e| \le |E_k(w^{mJ+j})e - E_k(w^{mJ})e| + |E_k(w^{mJ})e| \le \frac{C_6}{m} + |E_k(w^{mJ})e| \tag{24}$$

By (23), we can easily get

$$\lim_{m\to\infty}|E_k(w^{mJ+j})\cdot e| = 0,\ 1\le k\le N,\ 1\le j\le J \tag{25}$$

Because e is an arbitrary unit vector in $\mathbb{R}^P$, the weak convergence is obtained.

From (4) and (18), we have that for $1 \le k \le N,\ 1 \le j \le J$,

$$\lim_{m\to\infty} \|w_k^{mJ+j} - w_k^{mJ+j-1}\| = \lim_{m\to\infty} \|\triangle_j w_k^{mJ+j-1}\| \le \lim_{m\to\infty} \frac{C_5}{m} = 0 \qquad (26)$$

Thus

$$\lim_{i\to\infty} \|w^{i+1} - w^i\| = \lim_{i\to\infty} \|\big((w_1^{i+1} - w_1^i)^T, \cdots, (w_N^{i+1} - w_N^i)^T\big)^T\| = 0 \qquad (27)$$

So, this together with $(A3)$ and Lemma 2 leads to the strong convergence result $\lim_{i\to\infty} w^i = w^*$ and completes the proof.

References

1. Finnoff, W.: Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima, Neural Computation, **6** (1994) 285-295.
2. Hussaina, A.J. and Liatsisb, P.: Recurrent pi-sigma networks for DPCM image coding. Neurocomputing, **55** (2002) 363-382.
3. Li, Z.X., Wu, W. and Tian, Y.L.: Convergence of an online gradient method for feedforward neural networks with stochastic inputs. Journal of Computational and Applied Mathematics, **163** (2004) 165–176.
4. Shin, Y. and Ghosh, J.: The pi-sigma network: an efficient higher-order neural network for pattern classification and function approximation. International Joint Conference on Neural Networks, **1** (1991) 13-18.
5. Sinha, M., Kumar, K. and Kalra, P.K.: Some new neural network architectures with improved learning schemes. Soft Computing, **4** (2000) 214-223.
6. Xiong, Y., Wu, W., Lu, H.F., Zhang, C.: Convergence of online gradient algorithm method for Pi-Sigma neural Networks, to appear in Journal of Computational Information Systems.