Convergence of Online Gradient Method for Pi-sigma Neural Networks with Inner-penalty Terms

Kh. Sh. Mohamed¹, ², Xiong Yan³, Y. Sh. Mohammed⁴, ⁵, Abd-Elmoniem A. Elzain⁵, ⁶, Habtamu Z. A.², Abdhraman M. Adam²

¹Mathematical Department, College of Science, Dalanj University, Dalanj, Sudan
²School of Mathematical Sciences, Dalian University of Technology, Dalian, China
³School of Science, Liaoning University of Science & Technology, Anshan, China
⁴Physics Department, College of Education, Dalanj University, Dalanj, Sudan
⁵Department of Physics, College of Science & Art, Qassim University, Oklat Al- Skoor, Saudi Arabia
⁶Department Department of Physics, University of Kassala, Kassala, Sudan

Email address: khshm7@yahoo.com (Kh. Sh. Mohamed), xy-zhxw@163.com (Xiong Yan), yshm@yahoo.com (Y. Sh. Mohammed), Abdelmoniem1@yahoo.com (Abd-Elmoniem A. E.), habtamuz@mail.dlut.edu.cn(Habbitmu Z. A), abdelrhaman013@yahoo.com(A. M. Adam)

To cite this article:

Received: March 14, 2016; Accepted: March 30, 2016; Published: May 10, 2016

Abstract: This paper investigates an online gradient method with inner-penalty for a novel feed forward network it is called pi-sigma network. This network utilizes product cells as the output units to indirectly incorporate the capabilities of higher-order networks while using a fewer number of weights and processing units. Penalty term methods have been widely used to improve the generalization performance of feed forward neural networks and to control the magnitude of the network weights. The monotonicity of the error function and weight boundedness with inner-penalty term and both weak and strong convergence theorems in the training iteration are proved.

Keyword: Convergence, Pi-sigma Network, Online Gradient Method, Inner-penalty, Boundedness

1. Introduction

A novel higher order feedforward polynomial neural network is known to provide inherently more powerful mapping abilities than traditional feed forward neural network called the pi-sigma network (PSN) [2]. This network utilizes product cells as the output units to indirectly incorporate the capabilities of higher-order networks while using a fewer number of weights and processing units. The neural networks consisting of the PSN modules has been used effectively in pattern classification and approximation problems [1, 4, 10, 11]. There are two ways of training to updating weight: The first approach, batch (offline) training[18], the weights are modified after each training pattern is presented to the network. Second approach, online training, the weights updating immediately after each training sample is fed see [13]. The penalty term is often introduced into the network training algorithms has been widely used so as to control the magnitude of the weights and to improve the generalization performance of the network [6, 8], here the generalization performance refers to the capacity of a neural network to give correct outputs for untrained data. Specially cause, in the second approach the training weights updating become very large and over-fitting tends to occur, by adding the penalty term into the cost function, when use second approach has been successfully application see [3, 7, 12, 14], which acts as a brute-force to drive unnecessary weights to zero and to prevent the weights from taking too large in the training process. In the work area of penalty term at the same of the inner-penalty term (IP), which have worked to reduce the magnitude of the network weights with efficiency
improve the generalization performance of the network [5, 9, 17]. In this paper, we prove the (strong and weak) convergence of the online gradient with inner penalty and the monotonicity of the error function and the weight sequence are uniformly bounded during the training procedure with inner-penalty.

The rest of this paper is organized as follows. The neural network structure and the online gradient method with inner-penalty are described in Section 2. The preliminary lemmas are disruption in Section 3. The convergence results are presented and the rigorous proofs of the main results are provided in Section 4. Finally, in Section 5 we conclusions this study.

2. PSN-IPAlgorithm

PSN is a higher order feed forward polynomial neural network consisting of a single hidden layer. The hidden layer has summing units where the output layer has product units. PSN, which has a three-layer network consisting of input units, summing units, and output units, are described in Section 2. The preliminary lemmas and convergence results are presented and the rigorous proofs of the main results are provided in Section 4. Finally, in Section 5 we conclusions this study.

2. PSN-IPAlgorithm

PSN is a higher order feed forward polynomial neural network consisting of a single hidden layer. The hidden layer has summing units where the output layer has product units. PSN, which has a three-layer network consisting of input units, summing units, and 1 product layers. Let \(b_k = (b_{k1}, b_{k2}, ..., b_{kn})^T \in \mathbb{R}^n \) (1 \(\leq k \leq N \)) the weight vectors connecting the input and summing units, and write \(\omega = (b_1, b_2, ..., b_n) \in \mathbb{R}^{nN} \). We have included a special input unit \(\xi_p \), corresponding to the biases \(b_{kp} \), with fixed value-1. The structure of PSN is shown in Figure 1.

The network supplied with a given set of training samples \((\xi^j, \alpha^j)_{j=1}^N \in \mathbb{R}^p \times \mathbb{R} \). The error function with a inner penalty given by

\[
E(\omega) = \frac{1}{2} \sum_{j=1}^{N} (o^j - g(\prod_{i=1}^{N} (\alpha_i \cdot \xi^j)))^2 + \sum_{k=1}^{N} (\omega_k \cdot \xi^j)^2
\]

Where \(\lambda > 0 \) is a inner penalty coefficient and \(g(t) = \frac{1}{2}(o^j - g(t))^2 \) The gradient function is given by

\[
E_{\omega_k}(\omega) = \sum_{j=1}^{N} |g_j(\prod_{i=1}^{N} (\alpha_i \cdot \xi^j))| \prod_{i=k}^{N} (\alpha_i \cdot \xi^j) + \lambda (\omega_k \cdot \xi^j)\xi^j
\]

Finally, in Section 5 we conclusions this study.

3. Preliminary Lemmas

The next lemma present the monotonicity of the sequence \(\{E(\omega)\} \). It is essential for the proof of weakly convergence of PSN with penalty, presented in the following Theorems. For sake of description, we denote

\[
r_{k,j}^m = \Delta \omega_k^m + \Delta \omega_k^m
\]

\[
\psi_j^m = \prod_{i=1}^{N} (\alpha_i^m \cdot \xi^j)
\]

\[
\varphi_{k,j}^m = \prod_{i=k}^{N} (\alpha_i^m \cdot \xi^j)
\]

\[
1 \leq j \leq J, 1 \leq i \leq N, m = 0,1,...
\]

To begin with, first we present a few lemmas as preparation to prove Theorems

**Lemma 1.** Let Assumption 1–2 are valid, there hold

\[ |\psi_j^m - \psi_j^m| \leq M_2 \sum_{k=1}^{N} |\Delta \omega_k^m| \]

\[ |\varphi_{k,j}^m - \varphi_{k,j}^m| \leq M_2 \sum_{k=1}^{N} |\Delta \omega_k^m| \]

**Proof.** By Assumption 2 and Cauchy-Schwarz inequality, we have

\[
|\psi_j^m - \psi_j^m| \leq \sum_{j=1}^{N-1} |(\alpha_i^m \cdot \xi^j)| |(\alpha_i^m - \omega_k^m)\xi^j|
\]
Taylor’s formula we obtain
\begin{align}
& \sum_{l=1}^{N} (\omega_{l}^{m+1} \cdot \xi^{j}) (\omega_{N}^{m} \cdot \xi^{j}) (\omega_{N-1}^{m+1} - \omega_{N-2}^{m+1}) \xi^{j} \\
& + \cdots + \sum_{l=2}^{N} (\omega_{l}^{m+1} \cdot \xi^{j}) (\omega_{l-1}^{m+1} - \omega_{l-2}^{m+1}) \xi^{j} \\
& \leq M_{N}^{\frac{1}{2}} \sum_{k=1}^{N} \sum_{k=1}^{N} |\Delta \omega_{k}^{m+1}| \\
& \leq M_{2} \sum_{k=1}^{N} |\Delta \omega_{k}^{m+1}| 
\end{align}

where \( M_{2} = M_{N}^{\frac{1}{2}} \). Similarly, we get
\begin{align}
|\phi_{k,j,l}^{m+1} - \phi_{k,j,l}^{m}| & \leq M_{2} \sum_{k=1}^{N} |\Delta \omega_{k}^{m+1}| \tag{12}
\end{align}

Lemma 2 Suppose Assumptions 1–3 are satisfied, and the weight sequence \( \omega_{k}^{m} \) is generated by (4) – (5), then
\begin{align}
E(\omega_{k}^{m+1}) - E(\omega_{k}^{m}) & \leq - \left( \frac{1}{\eta} - \lambda - M_{3} - M_{4} - M_{5} \right) \sum_{k=1}^{N} |\Delta \omega_{k}^{m+1}| \| \xi \|^{2} \tag{13}
\end{align}

Proof. Applying Taylor’s formula to extend \( g_{i}(\psi_{j}^{m+1}) \) at \( (\psi_{j}^{m}) \), we have
\begin{align}
g_{i}(\psi_{j}^{m+1}) - g_{i}(\psi_{j}^{m}) &= g_{i}(\psi_{j}^{m}) (\phi_{k,j,l}^{m+1} - \phi_{k,j,l}^{m}) + \frac{1}{2} g''(t_{j})(\psi_{j}^{m+1} - \psi_{j}^{m})^{2} \\
& + \frac{1}{2} \sum_{k=1}^{N} (\omega_{k}^{m+1} - \omega_{k}^{m})^{2} \tag{14}
\end{align}

where \( t_{j} \in \mathbb{R} \) are on the line segment between \( \psi_{j}^{m+1} \) and \( \psi_{j}^{m} \). After dealing with (14) by accumulation \( g_{i}(\psi_{j}^{m+1}) \) for \( 1 \leq j \leq J \), we obtain from (2), (4), (5) and Taylor’s formula we obtain
\begin{align}
E(\omega_{k}^{m+1}) - E(\omega_{k}^{m}) &= \sum_{j=1}^{J} [g_{i}(\psi_{j}^{m+1}) - g_{i}(\psi_{j}^{m})] \\
& + \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{N} (\omega_{k}^{m+1} - \omega_{k}^{m})^{2} \tag{15}
\end{align}

where
\begin{align}
& |\mathcal{Z}_{1}| \leq \frac{1}{\eta} \sum_{k=1}^{N} (\Delta \omega_{k}^{m}) \cdot \sum_{k=1}^{N} (\Delta \omega_{k}^{m}) \tag{16} \\
& |\mathcal{Z}_{2}| \leq \frac{1}{2} \sum_{j=1}^{J} g''(t_{j})(\psi_{j}^{m+1} - \psi_{j}^{m})^{2} \tag{17} \\
& |\mathcal{Z}_{3}| \leq \frac{1}{2} \sum_{k=1}^{N} (\omega_{k}^{m+1} - \omega_{k}^{m})^{2} \tag{18} \\
& |\mathcal{Z}_{4}| \leq \sum_{k=1}^{N} \sum_{k=1}^{N} |\Delta \omega_{k}^{m+1}| \| \xi \|^{2} \tag{19}
\end{align}
4. Convergence Theorems

Now, we can elucidate and proofs the convergence theorems, which we needed.

**Theorem 1.** (Monotonicity): Let Assumption 1–3 are valid and the weight sequence \( \{ \omega_k^m \}_{m=0,1,\ldots} \) be generated by (4) ~ (5), then

\[
E(\omega_k^{m+1}) - E(\omega_k^m), \ m = 0, 1, \ldots
\]

**Proof.** Let

\[
M = M_3 + M_4 + M_5
\]

By Assumption 3, which satisfies

\[
0 < \eta < \frac{1}{\lambda M + M}
\]

Thus with (28) and Lemma 2, we have

\[
E(\omega_k^{m+1}) - E(\omega_k^m) \leq -\left( \frac{1}{\eta} - \lambda \bar{M} - M \right) \sum_{k=1}^{\infty} \| \Delta \omega_k^m \|^2 \leq 0
\]

This completes the proof of the Theorem 1.

**Theorem 2.** (Boundedness): Suppose that Assumption of Theorem 1 are valid, the weight sequence \( \{ \omega_k^m \}_{m=0,1,\ldots} \) be generated by (4) ~ (5) are uniformly bounded.

**Proof.** By Assumption 1, and Theorem 1, we have

\[
E(\omega_k^m) - E(\omega_k^{m-1}) \leq \cdots \leq E(\omega_0^m)
\]

\[
= \sum_{j=1}^{J} E(\psi_j^0) + \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{\infty} (\omega_k^0 \cdot \xi_j)^2 \leq C
\]

and

\[
C = J M_1 \left( 1 + \frac{1}{2} \sum_{k=1}^{\infty} \| \omega_k^0 \|^2 \right)
\]

From (2), (30) gives

\[
\lambda (\omega_k^0 \cdot \xi_j^0)^2 \leq 2E(\omega_k^m) \leq 2C , \ j = 1, 2, \ldots, J
\]

By (4) ~ (5), we have

\[
\omega_k^m = \omega_0^m - \eta \sum_{t=1}^{m} \gamma_{t} \psi_{t}^{0} \psi_{t}^{j} \xi_{k}^j + \lambda (\omega_k^0 \cdot \xi_j^0)^2
\]

Let the second part of above equation be \( \omega_k^m.\) Denote \( \omega_k^0 = \psi_0 \) and \( \omega_k^0 = \xi_1 \) be generated by (4) ~ (5), then

\[
\omega_k^0 = \omega_k^0 + \omega_k^2, \ \text{where} \ \omega_k^0 \in \mathbb{R}_1, \ \text{and} \ \omega_k^0 \in \mathbb{R}_2. \ \text{Then} \ \omega_k^m = (\omega_k^0 + \omega_k^2) \oplus \omega_k^2 = \omega_k^0 \oplus \omega_k^2. \ \text{Applying this to (33) we have}
\]

\[
|d_t| := |\omega_k^0 \cdot \xi_j^0| = |\omega_k^m \cdot \xi_j^0| \leq \frac{2\xi_j}{\sqrt{\lambda}}, \ t = 1, 2, \ldots, T
\]

Suppose \( \{ \xi_j^1, \xi_j^2, \ldots, \xi_j^T \} \in \{ 1, \ldots, \} \) is a base of the space \( \mathbb{R}_J. \) There are \( \alpha_t = \mathbb{R}_1 \in \{ 1, 2, \ldots, \} \) such that \( \omega_k^m = \alpha_t \xi_j^1 + \cdots + \alpha_T \xi_j^T. \) Then

\[
\left( \theta^t \xi_j^1 + \cdots + \theta^T \xi_j^T \right)^{t} = \left( d_{1}^{t} \right)^{t} \leq \frac{2\xi_j}{\sqrt{\lambda}}, \ t = 1, 2, \ldots, T
\]

Is a base, the coefficient determinant equal to zero, and the system of the linear equations has a unique solution. Assume that the coefficient determinant equals to. Then the solution is as follows

\[
L = \begin{bmatrix}
\xi_j^1 & \xi_j^2 & \cdots & \xi_j^T \\
\vdots & \vdots & \ddots & \vdots \\
\xi_j^1 & \xi_j^2 & \cdots & \xi_j^T \\
\xi_j^1 & \xi_j^2 & \cdots & \xi_j^T \\
\end{bmatrix}
\]

Then the solution is as follows

\[
a_t = L \cdot D^{-1}
\]

Let the maximum absolute value of all the sub-determinant with rank \( (T - 1) \) of the coefficient determinant is \( D^T, \) then

\[
|d_t| \leq |\theta^T| \cdot D^{-1} \cdot \sum_{i=0}^{T} d_i. \ \text{By (34) we have}
\]

\[
|d_t| \cdot T \cdot \frac{2\xi_j}{\sqrt{\lambda}}, \ t = 1, 2, \ldots, T. \ \text{Denote} \ \hat{C} = \max_{1 \leq t \leq T}\left\| \xi_j^t \right\|, \ \text{then}
\]

\[
\| \omega_k^m \| = \left\| \alpha_t \xi_j^1 + \cdots + \alpha_T \xi_j^T \right\| \leq |D^T \cdot D^{-1} \cdot \hat{C} \cdot T^2 \cdot \frac{2\xi_j}{\sqrt{\lambda}}
\]

That is \( \omega_k^m \) is bounded uniformly bounded. So from (29), we know \( \omega_k^m \) is uniformly bounded. In all, we get \( \{ \omega_k^m \}_{m=0,1,\ldots} \) are uniformly bounded, i.e., there exist a bounded closed region \( D \subset \mathbb{R}^k \) such that \( \{ \omega_k^m \} \subset D. \)

**Theorem 3.** (Weak convergence): Suppose that Assumption 1–3 are valid and the weight sequence \( \{ \omega_k^m \}_{m=0,1,\ldots} \) be generated by (4) ~ (5), then

\[
\lim_{m \to \infty} \| E_{\omega_k}(\omega_k^m) \| = 0.
\]

Furthermore, if Assumption 4 is also valid, we have the strong convergence: There exists \( \omega^* \in \Theta_0 \) such that

\[
\lim_{m \to \infty} \omega_k^m = \omega^*
\]
Proof. By (28) and setting $\beta > 0$ such that $\beta = \frac{1}{\eta} - \lambda M - M$, we have
\[
E(\omega_k^{m+1}) \leq E(\omega_k^m) - \beta \sum_{k'=1}^N \| \Delta \omega_k^m \|^2 \\
\leq \cdots \leq E(\omega_0^0) - \beta \sum_{k'=0}^m \sum_{k=1}^N \| \Delta \omega_k^m \|^2
\]
(40)
Since $E(\omega_k^{m+1}) > 0$ for any $m = 0,1,\ldots$. We set $m \to \infty$
\[
\sum_{k'=0}^m \sum_{k=1}^N \| \Delta \omega_k^m \|^2 \leq \frac{g(\omega_0^0)}{\beta} < \infty
\]
(41)
Combining (3) ~ (5), immediately gives the weak convergence result:
\[
\lim_{m \to \infty} \| E(\omega_k^m) \| = 0, \ m = 0,1,\cdots
\]
(42)
Next we prove the strong convergence it follows from (4)~(5) and (42) that leads
\[
\lim_{m \to \infty} \| \Delta \omega_k^m \| = 0, \ 0 \leq k \leq N
\]
(43)
Note that the error function $E(\omega^m)$ defined in (2) is continuously differentiable. By (43), Assumptions 4 and Lemma 3, immediately get the desired result. This completes the proof.

5. Conclusion

Through our study of this paper, the monotonicity of the error function $E(\omega^m)$ in formula (2) and the weight sequence boundedness $(\omega_k^m)_{m=0,1,\cdots}$ via formula (4) ~ (5) for the online gradient method with inner-penalty are presented, under those condition both weakly and strongly convergence theorems are proved.

Acknowledgment

We gratefully acknowledge Dalanj University for supporting this research. And our thanks to the anonymous reviewers and the editors for their previous helpful comments and valuable suggestion for their kind helps during the period of the research.

References


[3] P L Bartlett, For valid generalization, the size of the weights is more important than the size of the network, Advances in Neural Information Processing Systems 9 (1997) 134–140.


