Convergence of Offline Gradient Method with Inner-penalty for Multi-output Feedforward Neural Networks

Fengqi Zhou [1] Ergen Liu Yu Xiao

Abstract. In this paper, we study an offline gradient method with inner-penalty for training multi-output feedforward neural networks. The monotonicity of the error function and weight boundedness for the offline gradient with inner-penalty are presented, both weak and strong convergence results are proved, which will be very meaningful for theoretical research or applications on multi-output neural networks.

Keywords: Feedforward neural networks; Offline gradient method; Inner-penalty; Convergence

1.1 Introduction

Feedforward neural networks have been widely used in many applications [1-5]. The generalization ability is very essential for network performance, the generalization capability refers to the ratios of correctly classified untrained samples. A rule of thumb for improving the generalization is to choose the smallest network that fit the training examples. However, a simple but efficient way to restrict weight magnitude is to add some penalty terms to error function. So punishing term methods are often introduce into the neural networks training process and have proved to reduce the magnitude of the network weights and to effectively improve the generalization capability of the network [6-8].

The convergence of gradient method for the network training with one output unit has been considered by many authors [9-14]. The convergence of the online and batch gradient algorithm with a penalty term for feedforward neural network

[1] Fengqi. Zhou (✉)
College of Basic Sciences, East China Jiaotong University, Nanchang 330013, China
This work was funded by the National Natural Science Foundation of China (No. 11164007) , the Natural Science Foundation of Jiangxi province (No. 20132BAB212007), the Scientific Project of Jiangxi Education Department of China (GJJ11107).
e-mail: Zhoufengqi2004@163.com

has been also discussed [9,12,15-17]. In addition, multi-output feedforward neural network is widely used in classification problems. The experimental results show that the out representation is also important for the network performance. So the convergence of multi-output neural network is very meaningful. In this paper, we study a multi-output BP neural network with inner-penalty and define a relation formula between the penalty parameter and the learning rate parameter, then use it to prove the weak and strong convergences of the offline gradient algorithm with inner-penalty. Additionally, the boundness of the new error function with inner-penalty is also guaranteed.

1.2 Network Struction and Learning Method with Inner-penalty

In this section, we consider a two-layer network consisting of *P* input nodes, *T* output nodes. Fig.1.1 illustrates the structure of a two-layer multi-output feed-forward neural network.

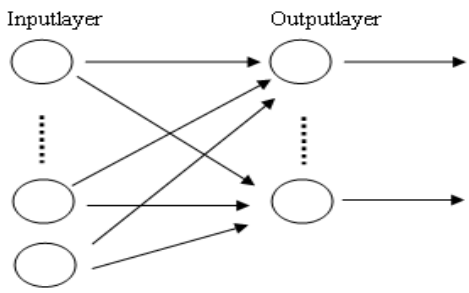

Fig. 1.1 Multi-output feedforward neural network.

Denote the weight matrix by $W = (w_{ij})_{TP}$ and $w_t = (w_{t1}, w_{t2}, \cdots, w_{tP})$. Assume that the transfer function $g : R \to R$ is a sigmoid function. Suppose $\{\xi^j, O^j\}_{j=1}^J$ is the given set of training examples.

Our error function with a penalty term has the following form

$$\begin{aligned} E(W) &= \frac{1}{2}\sum_{j=1}^{J}\sum_{t=1}^{T}[(O^j - g(w_t \cdot \xi^j))^2 + \mu(w_t \cdot \xi^j)^2] \\ &= \sum_{j=1}^{J}\sum_{t=1}^{T}[g_{jt}(w_t \cdot \xi^j) + \frac{1}{2}\mu(w_t \cdot \xi^j)^2] \end{aligned} \tag{2.1}$$

Where $\mu > 0$ is a penalty coefficient. Then gradient function is given by

$$E_{w_t}(W) = \sum_{j=1}^{J}[g'_{jt}(w_t \cdot \xi^j) + \mu(w_t \cdot \xi^j)]\xi^j \tag{2.2}$$

Let W^0 be arbitrary initial weights. We proceed to refine it iteratively by the following rule

$$w_t^{n+1} = w_t^n + \Delta w_t^n \qquad \Delta w_t^n = -\eta E_{w_t}(W) \tag{2.3}$$

Here, where the learning rate $\eta > 0$ is a constant. The following assumptions are needed for our boundedness and convergence results.

(A1): $\left|g^{(k)}(x)\right| \le C_3$, $\left|g_{jt}^{(k)}(x)\right| \le C_3$ $(k = 0,1,2)$ $x \in R$

(A2): μ and η are chosen to satisfy $0 \le \eta \le 1/\mu C + C_1$. Where

$$C = JC_2^2/2, C_1 = JC_3C_2^2/2, C_2 = \max_{1\le j\le J}\left\|\xi^j\right\|$$

1.3 Preliminary Theorem

Theorem 1: Suppose that assumptions A1), A2) hold. that the weight sequence $\{W^n\}$ is generated by the algorithm (2.3) for any initial value W^0 .Then we have

(*a*) $E(W^{n+1}) \le E(W^n) \qquad n = 0,1,2,\cdots;$

(*b*) *There is* $E^* \ge 0$ *such that* $\lim_{n\to\infty} E(W^n) = E^*$.

Proof: By the Taylor expansion and the learn rule (2.3)

$$E(W^{n+1}) - E(W^n) = \sum_{j=1}^{J}\sum_{t=1}^{T}(g_{jt}(w_t^{n+1}\cdot\xi^j) - g_{jt}(w_t^n\cdot\xi^j)) + \frac{\mu}{2}\sum_{j=1}^{J}\sum_{t=1}^{T}((w_t^{n+1}\cdot\xi^j)^2 - (w_t^n\cdot\xi^j)^2)$$

$$= \sum_{j=1}^{J}\sum_{t=1}^{T}[g'_{jt}(w_t^n\cdot\xi^j) + \mu(w_t^n\cdot\xi^j)]\Delta w_t^n\cdot\xi^j + \frac{1}{2}\sum_{j=1}^{J}\sum_{t=1}^{T}g''_{jt}(v_{t,j})(\Delta w_t^n\cdot\xi^j)^2 + \sum_{j=1}^{J}\sum_{t=}^{T}\frac{\mu}{2}(\Delta w_t^n\cdot\xi^j)^2$$

$$= -\frac{1}{\eta}\sum_{t=1}^{T}\left\|\Delta w_t^n\right\|^2 + (\mu C + C_1)\sum_{t=1}^{T}\left\|\Delta w_t^n\right\|^2$$

$$= -(\frac{1}{\eta} - \mu C - C_1)\sum_{t=1}^{T}\left\|\Delta w_t^n\right\|^2 \tag{3.1}$$

Here， $C = JC_2^2/2, C_1 = JC_3C_2^2/2$ 。

By assume A2, we have

$$E(W^{n+1}) \le E(W^n) \qquad n = 0,1,2,\cdots \tag{3.2}$$

Since the nonnegative sequence $E(W^n)$ is monotone and bounded below, there must be a limit value $E^* \ge 0$ such that $\lim_{n\to\infty} E(W^n) = E^*$.

Theorem 2 Suppose that Assumptions (A1) ,(A2) are valid. that the weight sequence $\{W^n\}$ is generated by the algorithm (2.3) for arbitrary initial value W^0. Then we have $\left\|w_t^n\right\|$ $(t=1,2,\cdots T; n=1,2,\cdots)$ are uniformly bounded, i.e., there exist a bounded closed region $\Phi \subset R^m$ such that $\{w_t^n\} \subset \Phi$.

Proof: By (3.2)

$$E(W^n) \le E(W^{n-1}) \le \cdots \le E(W^0) = \sum_{j=1}^{J}\sum_{t=1}^{T}[g_{jt}(w_t^0 \cdot \xi^j) + \frac{1}{2}\mu(w_t^0 \cdot \xi^j)^2] \le M \tag{3.3}$$

where $M = JTC_3 + \frac{1}{2}\mu JC_2^2 \sum_{t=1}^{T}\left\|w_t^0\right\|^2$.

From (2.1) and (3.3) we get

$$\mu(w_t^n \cdot \xi^j)^2 \le 2M \qquad j=1,2,\cdots,J \tag{3.4}$$

By (2.3)

$$w_t^n = w_t^0 - \eta\sum_{k=1}^{n-1}\sum_{j=1}^{J}[(g_j'(w_t^k \cdot \xi^j) + \mu(w_t^k \cdot \xi^j)]\xi^j \tag{3.5}$$

Let the second part of above equatio be w_{t1}^n , Denote $R_1 = span\{\xi^1, \xi^2, \cdots, \xi^J\} \subset R^m$ and $R_2 = R_1^{\perp}$ is the orthogonal complement space of R_1, obviously $w_{t1}^n \in R_1$, we divide w_t^0 into $w_t^0 = w_{t1}^0 + w_{t2}^0$, here $w_{t1}^0 \in R_1, w_{t2}^0 \in R_2$. Then $w_t^n = (w_{t1}^0 + w_{t1}^n) \oplus w_{t2}^0 = \tilde{w}_{t1}^n \oplus w_{t2}^0$. Applying this to (3.4) we have $\mu(\tilde{w}_{t1}^n \cdot \xi^j)^2 \le 2M$, i.e.,

$$\left|d_k\right| = \left|w_t^n \cdot \xi^k\right| = \left|\tilde{w}_{t1}^n \cdot \xi^k\right| \le \sqrt{2M/\mu} \quad k=1,2,\cdots,K \tag{3.6}$$

Asumme $\{\xi^{i_1}, \xi^{i_2}, \cdots, \xi^{i_K}\}$ $i_k \in \{1,\cdots,J\}, k=1,\cdots,K$ is a base of the space R_1. There are $a_k \in R(k=1,\cdots,K)$ such that $\tilde{w}_{t1}^n = a_1\xi^{i_1} + \cdots + a_K\xi^{i_K}$. Then $(a_1\xi^{i_1} + \cdots + a_K\xi^{i_K}) \cdot \xi^{i_k} = d_k, k=1,2,\cdots,K$. We get

$$\begin{pmatrix} \xi^{j_1} \cdot \xi^{j_1} & \cdots & \xi^{j_K} \cdot \xi^{j_1} \\ \vdots & \vdots & \vdots \\ \xi^{j_1} \cdot \xi^{j_K} & \cdots & \xi^{j_K} \cdot \xi^{j_K} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_K \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_K \end{pmatrix} \tag{3.7}$$

Because $\{\xi_1, \xi_2, \cdots, \xi_K\}$ is a base, the coefficient determinant is not equal to zero, thus the system of linear equations have an unique solution. Suppose the coefficient determinant equals to D, then the solution is as follows:

$$a_k = \begin{vmatrix} \xi^{j_1}\cdot\xi^{j_1} & \cdots & \xi^{j_{k-1}}\cdot\xi^{j_1} & d_1 & \xi^{j_{k+1}}\cdot\xi^{j_1} & \cdots & \xi^{j_K}\cdot\xi^{j_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \xi^{j_1}\cdot\xi^{j_K} & \cdots & \xi^{j_{k-1}}\cdot\xi^{j_K} & d_K & \xi^{j_{k+1}}\cdot\xi^{j_K} & \cdots & \xi^{j_K}\cdot\xi^{j_K} \end{vmatrix} D^{-1} \quad (3.8)$$

Let the maximum absolute value of all the subdeterminant with rank(K-1) of the coeffient determinant is D', the $|a_k| \le |D^{-1}||D'|\sum_{j=1}^{K} d_j$. By $|d_k| \le \sqrt{2M/\mu}$ we have $|a_k| \le K|D^{-1}||D'|\sqrt{2M/\mu}$ $(k = 1,2,\cdots,K)$. By (A2), we deduce that

$$\|\tilde{w}_{t1}^n\| = \|a_1\xi^{i_1} + a_2\xi^{i_2} + \cdots + a_K\xi^{i_K}\| \le K^2 C_2 |D^{-1}||D'|\sqrt{2M/\mu} \quad (3.9)$$

That is $\tilde{w}_{t1}^n$ are uniformly bounded. So from (3.5), we know w_t^n are uniformly bounded. In all, we get $\{w_t^n\}_{n=0}^{\infty}$ are uniformly bounded, i.e. there exist a bounded closed region $\Phi \subset R^m$ such that $\{w_t^n\} \subset \Phi$.

Theorem 3 Suppose that Assumptions (A1) ,(A2) are valid, $\Phi_0 = \{W | \nabla E(W) = 0\}$ is a finite point sets, that the weight sequence $\{W^n\}$ is generated by the algorithm (2.3) for arbitrary initial value W^0, we have

(1) weak convergence theorem: $\lim_{n\to\infty}\|E_{w_t}(W^n)\| = 0 \quad n = 1,2,\cdots$

(2) strong convergence theorem: there exists $W^* \in \Phi_0$ such that $\lim_{n\to\infty} W^n = W^*$

proof: Let $\gamma = 1/\eta - \mu C - C_1$, by Assumptions (A2), we have $\gamma > 0$. In view of (3.1), there holds

$$E(W^{n+1}) \le E(W^n) - \gamma\sum_{t=1}^{T}\|\Delta w_t^n\|^2 \le \cdots \le E(W^0) - \gamma\sum_{k=0}^{n}\sum_{t=1}^{T}\|\Delta w_t^n\|^2 \quad (3.10)$$

Since $E(W^{n+1}) \ge 0$ for any $n = 0,1,2,\cdots$. We let $n \to \infty$

$$\sum_{n=0}^{\infty}\sum_{t=1}^{T}\|\Delta w_t^n\|^2 \le E(w^0)/\gamma < \infty \quad (3.11)$$

Combining (2.2.) and (2.3) gives

$$\lim_{n\to\infty}\|E_{w_t}(W^n)\| = 0 \quad n = 1,2,\cdots \quad (3.12)$$

Thus (3.12) imply

$$\lim_{n\to\infty}\|w_t^{n+1} - w_t^n\| = 0 \quad (3.13)$$

This together with (3.12) and (3.13) leads to: existing $W^* \in \Phi_0$ such that $\lim_{n\to\infty} W^n = W^*$. This completes the proof.

1.4 Conclusion

In summary, we study an offline gradient method with inner-penalty for training multi-output feedforward neural networks. The monotonicity of the error function and weight boundedness for the offline gradient with inner-penalty are presented, both weak and strong convergence results are proved, which will provides a strong theoretical support for many applications on multi-output neural networks.

1.5 References

1. Y.C.Liang et al (2002).,Successive approximation training algorithm for feedforward neural networks. Neurocomputing 42: 311-322.
2. D.E.Rumelhart, G.E.Hinton an dR.J.Williams (1986) Learning Internal Representations by Error Propagation, MIT Press:318-362.
3. Fengqi Zhou, Wei Wu (2007). Output Representation of BP Neural Networks for Classification Problems. Journal of Information and Computational Science. 4(1):413-420.
4. Wei Wu (2003). Neural network computing. Beijing: Higher Education Press,. (in chinese)
5. Jie Yang, Wenyu Yang, Wei Wu (2012). A remark on the error-backpropagation learning algorithm for spiking neural networks. Applied Mathematics Letters 25:1118–1120.
6. Kong J, Wu W (2001). Online gradient methods with a punishing term for neural networks. Northeast Math. J 173: 371-378.
7. Setiono R (1997). A penalty-funtion approach for pruning feedforward neural networks. Neural Computation. 9: 185-204
8. Reed R (1993). Pruning algorithms-a survey. IEEE Trans. On Neural Network 4(5):74-747
9. Wei Wu, Hongmei Shao, Zhengxue Li (2006). Convergence of batch BP algorithm with penalty for FNN training. Lecture Notes in Computer Science 4232:562-569
10. Dongpo Xu, ZhengXue Li and Wei Wu (2008) Convergence of approcimated graeient method for Elmam networ . Neural Network World 18(3): 171-180.
11. Hongmei Shao, Gaofeng Zheng (2011), Convergence analysis of a back-propagation algorithm with adaptive momentum . Neurocomputing 74: 749-752.
12. Huisheng Zhang, Wei Wu (2009) Boundedness and convergence of online gradient method with penalty for linear output feedforward neural networks. Neural Process Lett 29: 205-212.
13. Hongmei Shao, Gaofeng Zheng (2011) Boundedness and convergence of online gradient method with penalty and momentum. Neurocomputing 74: 765-770.
14. Wei Wu, Jian Wang, Mingsong Cheng, Zhengxue Li (2011). Convergence analysis of online gradient method for BP neural networks. Neural Networks 24: 91-98.
15. Hongmei Shao, Wei Wu, Feng Li (2005) Convergence of online gradient method with a penalty term for feedforward neural networks with stochastic inputs.Nemerical Mathematics 1(14): 87-96.

16. Jian Wang, Wei Wu, et al (2012) Computational properties and convergence analysis of BPNN for cyclic and almost cyclic learning with penalty. Neural Networks 33: 127-135.
17. Huisheng Zhang, Wei Wu (2012). Boundedness and convergence of batch back-propagation algorithm with penalty for feedforward neural networks. Neurocomputing 89: 141-146.