CONVERGENCE OF ONLINE GRADIENT METHOD WITH A PENALTY TERM FOR FEEDFORWARD NEURAL NETWORKS WITH STOCHASTIC INPUTS*

Shao Hongmei(邵红梅) Wu Wei(吴 微) Li Feng(李 峰)

Abstract *Online gradient algorithm has been widely used as a learning algorithm for feedforward neural network training. In this paper, we prove a weak convergence theorem of an online gradient algorithm with a penalty term, assuming that the training examples are input in a stochastic way. The monotonicity of the error function in the iteration and the boundedness of the weight are both guaranteed. We also present a numerical experiment to support our results.*

Key words *Feedforward neural network, Online gradient algorithm, Penalty term, Stochastic input, Convergence, Monotonicity, Boundedness.*

AMS(2000)subject classifications 68T15

1 Introduction

Online gradient algorithm (OGM) is commonly used for feedforward neural network (FNN) training [2,3,5,6]. The training is usually done by iteratively updating of the weights according to the error signal, which is the negative gradient of a sum-square error function (SSE). However, by using SSE as the error function sometimes the weight of the network becomes very large and the generalization performance is poor, even though the network is trained until the error on the training set is minimized. In order to resolve this problem, a popular choice is to add a penalty term to the standard error function [1,4,8,9]. When the training samples are trained

* Partly supported by the National Natural Science Foundation of China,and the Basic Research Program of the Committee of Science,Technology and Industry of National Defense of China.
Received: Sep. 15, 2004

in a fixed order, the effect of the penalty term in controlling the magnitude of the weight is investigated in [8,10]. However, the usual OGM chooses input ξ^i from the training samples $\{\xi^i, O^i\}$ in a stochastic order, which is important to help the training procedure to jump off from local minima. In this paper we shall show that, when input sample ξ^i is chosen in a specially stochastic order (cf.[7]), such an online gradient algorithm with a penalty term and stochastic inputs (POGM-S) is weakly convergent. Besides, the monotonicity of the error function in the training iteration and the boundedness of the weight are both guaranteed. We also present a simulation example to illustrate our results established in the paper. Experimental results indicate that, as well as being beneficial from controlling the magnitude of the weight, POGM-S makes the generalization performance of the network greatly improved.

For simplicity, a two-layer FNN is considered with N input nodes and one single output node. Assume that the transfer function $\sigma : \mathbb{R} \to \mathbb{R}$ is a pre-chosen sigmoid function, and denote the weight by $\omega = (w_1, \cdots, w_N)^T$. Suppose $\{\xi^i, O^i\}_{i=1}^J$ is the given set of training examples. Our error function with a penalty term has the form (cf.[8])

$$E(\omega) = \frac{1}{2}\sum_{i=1}^{J}\left(O^i - \sigma(\omega\cdot\xi^i)\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{J}(\omega\cdot\xi^i)^2 \equiv \sum_{i=1}^{J}[f_i(\omega\cdot\xi^i) + \frac{\lambda}{2}(\omega\cdot\xi^i)^2], \tag{1.1}$$

where $\lambda > 0$ is the coefficient of the penalty term. Then the gradient function is given by

$$\nabla E(\omega) = \sum_{i=1}^{J}[f_i^{'}(\omega\cdot\xi^i) + \lambda(\omega\cdot\xi^i)]\xi^i. \tag{1.2}$$

Now we introduce the POGM-S algorithm. Let $\{\xi^{n1}, \xi^{n2}, \cdots, \xi^{nJ}\}$ be a stochastic permutation of $\{\xi^1, \xi^2, \cdots, \xi^J\}$ in the n-th cycle of the training iteration. Starting from an initial value ω^0, we proceed to refine it iteratively by the following rule

$$\omega^{nJ+i} = \omega^{nJ+i-1} + \triangle_i^n \omega^{nJ+i-1}, \quad i = 1, 2, \cdots, J; n = 0, 1, \cdots, \tag{1.3a}$$

$$\triangle_i^n \omega^{nJ+i-1} = -\eta_n[f_{ni}^{'}(\omega^{nJ+i-1}\cdot\xi^{ni}) + \lambda(\omega^{nJ+i-1}\cdot\xi^{ni})]\xi^{ni}, \tag{1.3b}$$

where η_n is the learning rate in the n-th training cycle. For an initial value $\eta_0 > 0$, η_n changes after each cycle of training iteration according to

$$\frac{1}{\eta_n} = \frac{1}{\eta_{n-1}} + \beta, \qquad n = 1, 2, \cdots, \tag{1.4}$$

where $\beta > 0$ is a constant. The following assumption is imposed throughout the paper.

Assumption 1 There is $C > 0$ such that for any $t \in \mathbb{R}$ and $1 \leq i \leq J$

$$|f_i(t)| \leq C, \quad |f_i^{'}(t)| \leq C, \quad |f_i^{''}(t)| \leq C.$$