NUMERICAL MATHEMATICS, A Journal of Chinese Universities (English Series)
Numer. Math. J. Chinese Univ. (English Ser.), issue 3, vol. **16**: 193-202, 2007

Convergence of BP Algorithm for Training MLP with Linear Output

Hongmei Shao
(*College of Mathematics and Computational Science, China University of Petroleum, Dongying, 257061, China***)**
Wei Wu*
(*Department of Applied Mathematics, Dalian University of Technology, Dalian, 116023, China*
*E-mail: wuweiw@dlut.edu.cn***)**
Wenbin Liu
(*Institute of Computational Mathematics and Management Science, University of Kent, UK*
*E-mail: W.B.Liu@kent.ac.uk***)**

Received March 30, 2006; Accepted (in revised version) April 18, 2007

Abstract

The capability of multilayer perceptrons (MLPs) for approximating continuous functions with arbitrary accuracy has been demonstrated in the past decades. Back propagation (BP) algorithm is the most popular learning algorithm for training of MLPs. In this paper, a simple iteration formula is used to select the learning rate for each cycle of training procedure, and a convergence result is presented for the BP algorithm for training MLP with a hidden layer and a linear output unit. The monotonicity of the error function is also guaranteed during the training iteration.

Keywords: Multilayer perceptron; BP algorithm; Convergence; Monotonicity.
Mathematics subject classification: 92B20, 68T05

1. Introduction

MLPs have gained much popularity in recent years for the capability of successfully solving classification and approximation tasks. The commonly used two MLP neural network models are of the sigmoid type and sigmoid-linear type [6, 7, 11, 16]. The former refers to its characteristic of all the activation functions being sigmoid, mainly aiming at the solutions to classification problems. The latter means that, the activation functions for hidden layers are sigmoid while those for output layer are linear. Since the actual output of sigmoid-linear networks can be arbitrarily large or small, they have been applied in control and approximation theories with satisfactory results [2,4,7,11]. Convergence of algorithms for the sigmoid network training has been studied in the literature [3, 8, 12–14]. In this paper, we concentrate on the deterministic convergence analysis of BP algorithm for sigmoid-linear MLP network training, including weak convergence and strong convergence.

*Corresponding author.

In the learning process of a MLP model, BP algorithm remains to be a simple and popular systematic method for updating the weights of networks [6, 11]. BP algorithm is based on applying the steepest descent gradient approach to the minimization of a cost function representing the instantaneous error signals. Consider a single node in the output layer of such a network. Denote the cost function by $E(w)$, where w represents the weight vector. The general form of batch BP learning algorithm for a MLP network training can be described as follows

$$w^{n+1} = w^n - \eta_n E_w(w^n), \tag{1.1}$$

where w^n, η_n and $E_w(w^n)$ denote the weight value, the learning rate at the n-th iteration and the gradient value at the point w^n, respectively. This obviously can be viewed as a gradient approach for solving a particular optimization problem. In this paper, we concentrate on the deterministic convergence analysis of BP algorithm for sigmoid-linear MLP network training, under weaker assumptions than what the general optimization theory normally presumes. A typical procedure to guarantee the convergence of gradient algorithms for general optimization problems is line search [15]. Typical line search methods such as Armijo-Goldstein and Wolfe-Powell [1,5,15] require the learning rates to be bounded both up and below, and much increase computational work for large networks. We shall use a simple procedure to select the learning rates, which can guarantee the monotonicity of the error function and the convergence of the BP algorithm for training multilayer perceptions. This procedure does not involve line search and is defined in a simple recurrent manner, resulting in less computational effort.

In this paper, we choose the learning rate in each cycle of learning iteration by the rule

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

where $\beta > 0$ is a constant (see [12, 14], where an online gradient method is considered). For an arbitrary initial value $\eta_0 > 0$, there always hold $\eta_n \to 0$ as $n \to \infty$. In this way the oscillations can be avoided and the training time can be shortened, especially when the weights are moving along the wall of a ravine. One main work in this paper is to show that such a BP algorithm with the property $\eta_n = \mathcal{O}(\frac{1}{n})$ is convergent and the error sequence is descending in a monotonous way. The other is, as a remark, we extend the selection of learning rates to more general cases

$$\eta_n = \mathcal{O}(\frac{1}{n^\delta}), \quad \frac{1}{2} < \delta \leq 1,$$

and obtain similar convergence results.

The rest of this paper is organized as follows. The neural model and the batch BP learning algorithm with the learning rate $\eta_n = \mathcal{O}(\frac{1}{n})$ is described in the next section. Section 3 presents some lemmas and a convergence theorem. Their detailed proofs and the generalized results are gathered in Section 4. In this paper, we use in the proofs $\|\cdot\|$ for the Euclidean norm and C_i for genetic positive constants which are independent of the iteration n.