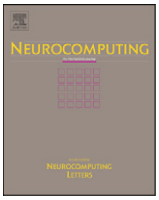
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



# Deterministic convergence of conjugate gradient method for feedforward neural networks ☆

Jian Wang [a,b,c], Wei Wu [a], Jacek M. Zurada [b,*]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China
[b] Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA
[c] School of Mathematics and Computational Sciences, China University of Petroleum, Dongying 257061, PR China

ARTICLE INFO

*Article history:*
Received 18 November 2010
Received in revised form
3 March 2011
Accepted 19 March 2011
Communicated by J. Cao
Available online 10 May 2011

*Keywords:*
Deterministic convergence
Conjugate gradient
Backpropagation
Feedforward neural networks

ABSTRACT

Conjugate gradient methods have many advantages in real numerical experiments, such as fast convergence and low memory requirements. This paper considers a class of conjugate gradient learning methods for backpropagation neural networks with three layers. We propose a new learning algorithm for almost cyclic learning of neural networks based on PRP conjugate gradient method. We then establish the deterministic convergence properties for three different learning modes, i.e., batch mode, cyclic and almost cyclic learning. The two deterministic convergence properties are weak and strong convergence that indicate that the gradient of the error function goes to zero and the weight sequence goes to a fixed point, respectively. It is shown that the deterministic convergence results are based on different learning modes and dependent on different selection strategies of learning rate. Illustrative numerical examples are given to support the theoretical analysis.

## 1. Introduction

The feedforward neural networks with backpropagation (BP) training procedure have been widely used in various fields of scientific research and engineering applications. The BP algorithm attempts to minimize the least squared error of objective function, defined by the differences between the actual network outputs and the desired outputs [1]. There are two popular ways of learning with training samples to implement the backpropagation algorithm: batch mode and incremental mode [2]. For batch training, weight changes are accumulated over an entire presentation of the training samples before being applied, while incremental training updates weights after the presentation of each training sample [3].

There are three incremental learning strategies according to the order that the samples are applied [2,4–6]. The first strategy is online learning (completely stochastic order). The second strategy is almost cyclic learning (special stochastic order) with each sample from the training set submitted once per cycle and with sample sequence fixed for each cycle. The last one is cyclic learning (fixed order).

In most cases, the feedforward neural networks are trained with supervised learning techniques which employ the steepest descent method [7,8]. We mention that the steepest descent method takes consecutive steps in the direction of negative gradient of the performance surface. There has been considerable research on the methods to accelerate the convergence of the steepest descent method [9,10]. Unfortunately, in practice, even with these modifications, the method exhibits oscillatory behavior when it encounters steep valleys, thus, it is not effective due to slow progress. An important reason for this is that the steepest descent method is a simple first order gradient descent method with poor convergence properties.

There have been a number of reports describing the use of second order numerical optimization methods to accelerate the convergence of backpropagation algorithm, such as conjugate gradient method (CG) and Newton method [8,11]. Newton method is much faster than the steepest descent method, but requires the Hessian matrix and its inverse to be calculated. The CG method is a kind of compromise algorithm; it does not require the calculation of second derivatives, and yet it still has the quadratic convergence property [12].

In general, conjugate gradient methods are much more effective than the steepest descent method and are almost as simple to compute. These methods do not attain the fast convergence rates of Newton or quasi-Newton methods, but they have the advantage of not requiring storage of matrices [13]. The linear conjugate gradient method was first proposed in [14] as an iterative method for solving linear systems with positive definite coefficient matrices. The first nonlinear conjugate gradient method was introduced in [15]. It is one of the earliest and most popular techniques for solving large scale nonlinear optimization problems. Different conjugate gradient methods have been proposed in recent years which depend on the different choices of the descent directions [16,17]. There are three

☆ Project supported by the National Natural Science Foundation of China (No. 10871220) and China Scholarship Council.

* Corresponding author. Tel.: +1 502 852 6314; fax: +1 502 852 3940.
*E-mail address:* jacek.zurada@louisville.edu (J.M. Zurada).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.03.016

classical CG methods such as FR [15], PRP [18,19] and HS [14] conjugate gradient methods. Among these methods, the PRP method is often regarded as the best one in practical computations [20].

The batch mode is commonly used for the PRP method in feedforward neural networks [21]. The cyclic learning for PRP method was first presented in [22]. The learning rate is adjusted automatically providing relatively fast convergence at early stages of adaptation while ensuring small final misadjustment for cases of stationary environments. For non-stationary environments, the cyclic learning method proposed in [22] has good tracking ability and quick adaptation to abrupt changes, as well as, to produce a small steady state misadjustment.

In [23], a novel algorithm is proposed for blind source separation based on the cyclic PRP conjugate gradient method. The line search method is applied to find the best learning rate. Simulations show the ability of the algorithm to perform the separation even with an ill-conditioned mixed matrix. To our best knowledge, the almost cyclic learning for the PRP method has not been discussed until now.

Convergence property for neural networks is an interesting research topic which offers an effective guarantee in practical applications. However, it is noted that there are other effective methods to train neural networks. A novel recurrent neural network based on the gradient method is proposed for solving linear programming problems in [24,25]. The finite time convergence is guaranteed by using the Lyapunov method. Furthermore, the network with simple structure converges globally to exact optimal solutions. As reported in [26], the extreme learning machine (ELM) algorithm based on the least-squares is more effective than gradient-based learning for feedforward neural networks in many applications [27,28]. An essential theoretical result in [26] is that single-hidden layer feedforward neural networks (SLFNs) with $N$ hidden nodes can learn $N$ distinct samples exactly and may require less than $N$ hidden nodes if learning error is allowed. The convergence property of SLFNs based on ELM algorithm is beyond the scope of this paper and is left for future investigation. In this paper, we just focus on the convergence property of feedforward neural networks based on conjugate gradient methods.

The convergence results for feedforward neural networks published in the literature mainly concentrate on the steepest descent method. Some weak and strong convergence results based on batch mode training process are proven with special assumptions in the recent paper [29]. The convergence results for online learning are mostly asymptotic convergence due to the arbitrariness in the presentation order of the training samples [30–33]. On the other hand, deterministic convergence lies in cyclic and almost cyclic learning mainly because every sample of the training set is fed exactly once in each training epoch [4–6,34,35].

In this paper, we present a novel study of the deterministic convergence of BP neural networks based on PRP method, including both weak and strong convergence. The weak convergence indicates that the gradient of the error function goes to zero, while the weight sequence itself goes to a unique fixed point for the strong convergence. We obtain the convergence conditions with a constant learning rate for batch mode, and a more general choice instead of line search for cyclic and almost cyclic learning. Specially, we demonstrate the following novel contributions:

(A) The almost cyclic learning of PRP conjugate gradient method is presented in this paper:

The almost cyclic learning is common for BP neural networks by employing the steepest descent method [35]. However, there is no report for almost cyclic learning BP neural networks based on PRP conjugate gradient method. We claim that the order of samples can be randomly arranged after each training cycle for almost cyclic PRP learning method.

(B) The deterministic convergence of batch mode conjugate gradient (BCG) is obtained which includes the strong convergence, that is, weight sequence goes to a fixed point:

For BCG method, we consider the case of a constant learning rate rule. The weak convergence for general nonlinear optimization problems is proved in [16]. However, we extend the convergence result including the strong convergence result as well in this paper. Xu et al. [29] prove the weak and strong convergence based on batch mode learning of steepest descent method for three complex-valued recurrent neural networks. We claim that the assumptions of the activation functions and the stationary points of error function in this paper are more relaxed than those in [29]. In addition, it is easy to extend the convergence results to the complex-valued recurrent neural networks. We mention that the BCG method would become the steepest descent method once the conjugate coefficients are set to zero.

(C) The deterministic convergence including weak convergence and strong convergence of cyclic conjugate gradient (CCG) for feedforward neural networks are obtained for the first time:

The PRP conjugate gradient method has no global convergence in many situations. Some modified PRP conjugate gradient methods with global convergence were proposed [36–39] via adding some strong assumptions or using complicated line searches. To our best knowledge, the deterministic convergence results in this paper are novel for CCG and ACCG (almost cyclic conjugate gradient) for feedforward neural networks. We note that cyclic learning with steepest descent method is a special case of CCG presented below. A dynamic learning strategy which depends on the instant conjugate direction is considered in [22]. However, from mathematical point of view, presented below is a more general case for learning rate instead of line search strategy.

(D) The above deterministic convergence results are also valid for ACCG.

Similarly, almost cyclic learning for steepest descent method is a special case of ACCG once the conjugate coefficients are set to zero in this paper. The convergence assumptions of ACCG for feedforward neural networks are more relaxed than those in [35].

The rest of this paper is organized as follows: In Section 2, three updating methods including BCG, CCG and ACCG are introduced. The main convergence results are presented in Section 3 and their proofs are in Section 5. Conclusions are drawn in Section 6.

## 2. Algorithms

### 2.1. Conjugate gradient methods

Consider an unconstrained minimization problem

$$\min f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, \tag{1}$$

where $\mathbb{R}^n$ denotes an $n$-dimensional Euclidean space and $f : \mathbb{R}^n \to \mathbb{R}^1$ is a continuously differentiable function.

Generally, a line search method takes the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}_k, \quad k = 0,1,\ldots, \tag{2}$$

where $\mathbf{d}_k$ is a descent direction of $f(\mathbf{x})$ at $\mathbf{x}^k$ and $\alpha_k$ is a step size. For convenience, we denote $\nabla f(\mathbf{x}^k)$ by $\mathbf{g}_k$, $f(\mathbf{x}^k)$ by $f_k$ and $\nabla^2 f(\mathbf{x}^k)$