

A Framework for the Development of Globally Convergent Adaptive Learning Rate Algorithms

G.D. MAGOULAS^{1,3}, V.P. PLAGIANAKOS^{2,3},
G.S. ANDROULAKIS^{2,3} AND M.N. VRAHATIS^{2,3,*}

1. Department of Informatics, University of Athens, GR-157.84 Athens, GREECE

2. Department of Mathematics, University of Patras, GR-261.10 Patras, GREECE

3. University of Patras Artificial Intelligence Research Center,
GR-261.10 Patras, GREECE

Abstract: In this paper we propose a framework for developing globally convergent batch training algorithms with adaptive learning rate. The proposed framework provides conditions under which global convergence is guaranteed for adaptive learning rate training algorithms. To this end, the learning rate is appropriately tuned along the given descent direction. Providing conditions regarding the search direction and the corresponding stepsize length this framework can also guarantee global convergence for training algorithms that use a different learning rate for each weight. To illustrate the effectiveness of the proposed approach on various training algorithms simulation results are provided.

Keywords and phrases: Global convergence, learning rate adaptation, batch training algorithms, steepest descent, feedforward neural networks.

1. Introduction

Supervised neural network training is a subject of considerable ongoing research and numerous algorithms have been proposed to this end. A common approach is to realize training by minimizing the network learning error, which is a measure of its performance and is usually based on the difference between the actual output vector of the network and the desired output vector. The rapid computation of a set of weights that minimizes this error is a rather difficult task since, in general, the number of network weights is high and the error function generates a complicated surface in the weight space, possessing multitudes of local minima and having broad flat regions adjoined with narrow steep ones that need to be searched to locate an "optimal" weight set.

In order to simplify the formulation of the equations throughout the paper a unified notation for the weights is adopted. Thus, for a Feedforward Neural Network (FNN) with a total of n weights, \mathbb{R}^n is the n -dimensional real space of column weight vectors w with components w_1, w_2, \dots, w_n and w^* is the optimal weight vector with components $w_1^*, w_2^*, \dots, w_n^*$; E is the batch error measure defined as the sum-of-squared-differences error function over the entire training set; $\partial_i E(w)$ denotes the partial derivative of $E(w)$ with respect to the i th variable w_i ; $\nabla E(w)$ defines the gradient vector of the sum-of-squared-differences error function E at w while $H = [H_{ij}]$ defines the Hessian $\nabla^2 E(w)$ of E at w .

The special case of the batch training is consistent with the theory of unconstrained optimization. In this case the minimization corresponds to updating the weights after each presentation of the entire training set, which is called an *epoch*, and requires that the sequence of weight vectors $\{w^k\}_{k=0}^{\infty}$, where k indicates epochs, converges to a set w^* that minimizes E .

The widely used batch Back-Propagation (BP) [23] is a first-order training algorithm, which minimizes the error function using the steepest descent method [8]:

$$w^{k+1} = w^k - \eta \nabla E(w^k), \quad (1)$$

* Corresponding author. Tel.: +30-61-99734; Fax: +30-61-992965 (Michael N. Vrahatis) E-mail:& URL addresses: vrahatis@math.upatras.fr —http://www.math.upatras.gr/~vrahatis

where the gradient vector is usually computed by the back-propagation of the error through the layers of the FNN (see [23]) and η is a heuristically chosen constant parameter, called learning rate. Appropriate learning rates help to avoid convergence to a saddle or maximum point. In practice, a small constant learning rate is chosen ($0 < \eta < 1$) in order to secure the convergence of the BP algorithm and to avoid oscillations in the directions where the error surface is steep. However, this approach considerably slows down the training process since, in general, a small learning rate may not be appropriate for all the portions of the error surface.

Our motivation in this paper is to provide general theoretical results and strategies that are applicable to guarantee the convergence of adaptive learning rate algorithms. The algorithms differ according to the information they need to modify the learning rate. In training algorithms with a *global* learning rate, the same rate is used to update all the weights in the FNN, while in algorithms with a *local* learning rate a different learning rate is used for each weight.

The paper is organized as follows. Section 2 provides an overview of adaptive learning rate BP algorithms. In Section 3 the issues of monotone decrease of the error function, as well as the notion of global convergence are introduced. Then, strategies for developing globally convergent modifications of adaptive learning rate algorithms are presented in Sections 4 and 5, while in Section 6 we present an application example to evaluate and compare various adaptive learning rate algorithms. Finally, the paper ends in Section 7 with some concluding remarks.

2. Adaptive learning rate algorithms

Several adaptive learning rate algorithms have been proposed to accelerate the training procedure. The following strategies are usually suggested:

- (i) start with a small learning rate and increase it exponentially if successive epochs reduce the error, or rapidly decrease it if a significant error increase occurs [2, 25],
- (ii) start with a small learning rate and increase it if successive epochs keep gradient direction fairly constant, or rapidly decrease it if the direction of the gradient varies greatly at each epoch [4] or
- (iii) for each weight an individual learning rate is given, which increases if the successive changes in the weights are in the same direction and decreases otherwise [10, 19, 21, 24].

Note that all the above mentioned strategies employ heuristic parameters in an attempt to enforce the monotone decrease of the learning error and to secure the converge of the training algorithm to a minimizer of E .

A different approach is based on Goldstein's and Armijo's work on steepest-descent and gradient methods. The method of Goldstein [9] requires the assumption that E is twice continuously differentiable on $\mathcal{S}(w^0)$, where $\mathcal{S}(w^0) = \{w : E(w) \leq E(w^0)\}$ is bounded, for some initial vector w^0 . It also requires that η is chosen to satisfy the relation $\sup \|H(w)\| \leq \eta^{-1} < \infty$ in some bounded region, where the relation $E(w) \leq E(w^0)$ holds. The k th iteration of an algorithm model that follows this approach consists of the following steps:

1. Choose η_0 to satisfy $\sup \|H(w)\| \leq \eta_0^{-1} < \infty$ and δ to satisfy $0 < \delta \leq \eta_0$.
2. Set $\eta^k = \eta$, where η is such that $\delta \leq \eta \leq (2\eta_0 - \delta)$ and go to the next step.
3. Update the weights $w^{k+1} = w^k - \eta^k \nabla E(w^k)$.

However, the manipulation of the full Hessian is too expensive in computation and storage for FNNs with several hundred weights [3]. Le Cun [11] proposed a technique, based on appropriate perturbations of the weights, for estimating on-line the principle eigenvalues and eigenvectors of the Hessian without

calculating the full matrix H . According to experiments reported in [11] the largest eigenvalue of the Hessian is mainly determined by the FNN architecture, the initial weights and by short-term low-order statistics of the training data. This technique could be used to determine η_0 , in Step 1 of the above algorithm, requiring additional presentations of the training set in the early training.

An alternative approach is based on the work of Armijo [1]. Following this approach, the value of the learning rate η is related to the value of the Lipschitz constant K , which depends on the morphology of the error surface. In this case, the BP algorithm takes the form:

$$w^{k+1} = w^k - \frac{1}{2K} \nabla E(w^k), \quad (2)$$

and converges to the point w^* which minimizes E (see [1] for conditions under which convergence occurs and a convergence proof). However, in practice neither the morphology of the error surface nor the value of K are known a priori. In [13] a local estimation of the Lipschitz constant has been proposed, as part of a learning rate adaptation strategy that provides increased rate of convergence through the Lipschitz constant estimation and guarantees the stability of the learning procedure.

3. Monotone decrease of the error function and global convergence

A training algorithm can be made globally convergent by determining the learning rate in such a way that the error is exactly minimized along the current search direction at each epoch, i.e. $E(w^{k+1}) < E(w^k)$. To this end, an iterative search, which is often expensive in terms of error function evaluations, is required. It must be noted that the above simple condition does not guarantee global convergence for general functions, i.e. converges to a local minimizer from any initial condition (see [5] for a general discussion on globally convergent methods).

The use of adaptive learning rate algorithms which enforce monotonic error reduction using inappropriate values for the critical heuristic learning parameters can considerably slow the rate of training, or even lead to divergence and to premature saturation [12, 22]. Moreover, using heuristics it is not possible to develop globally convergent training algorithms.

To alleviate this situation it is preferable to tune the learning rate, which is evaluated by an adaptive learning rate algorithm, so that the value of the error function is sufficiently decreased at each epoch, accompanied by a significant change in the value of w . A strategy of this kind consists in accepting a positive learning rate η^k along the search direction φ^k if it satisfies the *Wolfe conditions*:

$$E(w^k + \eta^k \varphi^k) - E(w^k) \leq \sigma_1 \eta^k \langle \nabla E(w^k), \varphi^k \rangle, \quad (3)$$

$$\langle \nabla E(w^k + \eta^k \varphi^k), \varphi^k \rangle \geq \sigma_2 \langle \nabla E(w^k), \varphi^k \rangle, \quad (4)$$

where $0 < \sigma_1 < \sigma_2 < 1$ and $\langle \cdot, \cdot \rangle$ stands for the usual inner product in \mathbb{R}^n . The first inequality ensures that the error is reduced sufficiently and the second prevents the learning rate from becoming too small. It can be shown that if φ^k is a descent direction, if E is continuously differentiable and if E is bounded below along the radius $\{w^k + \eta \varphi^k \mid \eta > 0\}$, then there always exist learning rate satisfying (3)-(4) [5, 16]. Relation (4) can be replaced by [5]:

$$E(w^k + \eta^k \varphi^k) - E(w^k) \geq \sigma_2 \eta^k \langle \nabla E(w^k), \varphi^k \rangle, \quad \sigma_2 \in (\sigma_1, 1). \quad (5)$$

An alternative strategy has been proposed in [20]. It is applicable to any descent direction φ^k and uses two parameters $\alpha, \beta \in (0, 1)$. Following this approach the learning rate is $\eta^k = \beta^{m_k}$, where $m_k \in \mathbb{Z}$ is any integer such that:

$$E(w^k + \beta^{m_k} \varphi^k) - E(w^k) \leq \beta^{m_k} \alpha \langle \nabla E(w^k), \varphi^k \rangle, \quad (6)$$

$$E(w^k + \beta^{m_k-1} \varphi^k) - E(w^k) > \beta^{m_k-1} \alpha \langle \nabla E(w^k), \varphi^k \rangle. \quad (7)$$

epoch), but is enforced simply by placing a lower bound on the acceptable values of the learning rate. This bound on the learning rate has the same theoretical effect as the condition (4) and ensures global convergence [5]. The value $q = 2$ is usually suggested in the literature [1] and indeed it was found to work without problems in the experiments (see [14]).

In this framework, an important theorem due to Wolfe [5] states that if E is bounded below, then the sequence $\{w^k\}_{k=0}^{\infty}$ generated by any algorithm that follows a descent direction φ^k whose angle θ_k with $-\nabla E(w^k)$ is such that:

$$\cos \theta_k = \frac{\langle -\nabla E(w^k), \varphi^k \rangle}{\|\nabla E(w^k)\| \|\varphi^k\|} > 0, \quad (8)$$

and satisfy the Wolfe's conditions, will also obey $\lim_{k \rightarrow \infty} \nabla f(w^k) = 0$ [5, 16].

Theorem 1 [5, 16, 27, 28]. *Suppose that the error function $E : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable on \mathbb{R}^n and assume that ∇E is Lipschitz continuous on \mathbb{R}^n . Then, given any $w^0 \in \mathbb{R}^n$, either E is unbounded below, or there exists a sequence $\{w^k\}_{k=0}^{\infty}$ obeying the Wolfe's conditions (3)-(4) and either:*