ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 1-6

A New Learning Rates Adaptation Strategy for the Resilient Propagation Algorithm

Aristoklis D. Anastasiadis[1], George D. Magoulas[1], Michael N. Vrahatis[2]

[1] School of Computer Science and Information Systems,
Birkbeck College, University of London,
Malet Street, London WC1E 7HX, United Kingdom
[2] University of Patras Artificial Intelligence Research Center (UPAIRC),
Department of Mathematics, University of Patras, GR–26110, Greece

Abstract. In this paper we propose an Rprop modification that builds on a mathematical framework for the convergence analysis to equip Rprop with a learning rates adaptation strategy that ensures the search direction is a descent one. Our analysis is supported by experiments illustrating how the new learning rates adaptation strategy works in the test cases to ameliorate the convergence behaviour of the Rprop. Empirical results indicate that the new modification provides benefits when compared against the Rprop and a modification proposed recently, the Improved Rprop.

Keywords: Supervised learning, feedforward neural networks, convergence analysis, global convergence property, Rprop, Improved Rprop.

1 Introduction

The Resilient Propagation (Rprop) algorithm is one of the most popular adaptive learning rates training algorithms [9]. It employs a sign-based scheme to eliminate harmful influences of derivatives' magnitude on the weight updates, and is eminently suitable for applications where the gradient is numerically estimated or the error is noisy [2]; it is easy to implement in hardware and is not susceptible to numerical problems [5]. The ideas behind Rprop have motivated the development of several variants with the aim to improve the convergence behavior and effectiveness of the original method. Thus hybrid learning schemes have been proposed to incorporate second derivative related information in Rprop, such as the *QRprop*, which approximates the second derivative by one–dimensional secant steps, and the *Diagonal Estimation Rprop–DERprop* [7], which directly computes the diagonal elements of the Hessian matrix. Also approaches inspired from global optimisation theory have been developed to equip Rprop with annealing strategies, such as the *Simulated Annealing Rprop–SARprop* and the *Restart mode Simulated Annealing Rprop–ReSARprop* [11] in order to escape

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 1-6

from shallow local minima. Recently, the *Improved Rprop–IRprop* algorithm [2], which applies a backtracking strategy (i.e. it decides whether to take back a step along a weight direction or not by means of a heuristic), has shown improved convergence speed when compared against existing Rprop variants, as well as other training methods. Relevant literature shows that Rprop-based learning schemes exhibit fast convergence in empirical evaluations, but usually require introducing or even fine tuning additional heuristics. For example, annealing schedules require heuristics for the acceptance probability and the visiting distribution, whilst second derivative methods employ heuristics in the various approximations of the second derivative. Moreover, literature shows a lack of theoretical results underpinning the Rprop modifications, particularly the first–order methods. This is not surprising as heuristics make difficult to guarantee converge to a local minimiser of the error function when adaptive learning rates for each weight are used in calculating the weight updates [2, 3, 6, 9]. This paper proposes a new Rprop-based learning scheme and presents a theoretical justification for its development. In the next section, the new algorithm and the corresponding theoretical result are presented. Then results on the experimental evaluation of the algorithm as well as comparisons with the original Rprop and the recently proposed IRprop are reported. The paper ends with concluding remarks.

2 A Modification of the Rprop

In our approach Rprop's convergence to a local minimiser is treated with principles from unconstrained minimisation theory. Suppose that (i) $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$ is the function to be minimized and f is bounded below in $\mathbb{R}^n$; (ii) f is continuously differentiable in a neighborhood $\mathcal{N}$ of the level set $\mathcal{L} = \{x : f(x) \leqslant f(x^0)\}$, and (iii) ∇f is Lipschitz continuous on $\mathbb{R}^n$ that is for any two points x and $y \in \mathbb{R}^n$, ∇f satisfies the Lipschitz condition $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|, \quad \forall x, y, \in \mathcal{N}$ where $L > 0$ denotes the Lipschitz constant, and x^0 is the starting point of the following iterative scheme

$$x^{k+1} = x^k + \tau^k d^k. \tag{1}$$

Convergence of the general iterative scheme (1), in which d^k is the search direction and $\tau^k > 0$ is a step–length, requires that the adopted search direction d^k satisfies the condition $g(x^k)^\top d^k < 0$, which guarantees that d^k is a descent direction of $f(x)$ at x^k. The step–length in (1) can be defined by means of a number of rules, such as the Armijo's rule [1], the Goldstein's rule [1], or the Wolfe's rule [13, 14], and guarantees the convergence in certain cases. For example, when the step–length is obtained through Wolfe's rule [13, 14]

$$f(x^k + \tau^k d^k) - f(x^k) \leqslant \sigma_1 \tau^k g(x^k)^\top d^k, \tag{2}$$

$$g(x^k + \tau^k d^k)^\top d^k \geqslant \sigma_2 g(x^k)^\top d^k, \tag{3}$$

where $0 < \sigma_1 < \sigma_2 < 1$, and g is the gradient, then a theorem by Wolfe [13, 14] is used to obtain convergence results. Moreover, the Wolfe's Theorem [1, 4]

suggests that if the cosine of the angle between the search direction d^k and $-g(x^k)$ is positive then $\lim_{k\to\infty} g(x^k) = 0$, which means that the sequence of gradients converges to zero. For an iterative scheme (1), $\lim_{k\to\infty} g(x^k) = 0$ is the best type of global convergence result that can be obtained (see [4] for a detailed discussion). Evidently, no guarantee is provided that (1) will converge to a global minimiser, x^*, but only that it possesses the global convergence property, [1, 4], to a local minimiser, i.e. "is designed to converge to a *local* minimizer of a nonlinear function, *from almost any starting point*" [1, p.5].

In batch training, E is bounded from below, since $E(w) \geqslant 0$. For a given training set and network architecture, if w^* exists such that $E(w^*) = 0$, then w^* is a global minimiser; otherwise, w with the smallest $E(w)$ value is considered a global minimiser. Also, when using *smooth enough* activations (the derivatives of at least order p are available and continuous), such as the well known hyperbolic tangent, the logistic activation function etc., the error E is also smooth enough.

Theorem. Suppose that (i)-(iii) are fulfilled. Then, for any $w^0 \in \mathbb{R}^n$ and any sequence $\{w^k\}_{k=0}^{\infty}$ generated by the Rprop's scheme

$$w^{k+1} = w^k - \tau^k \,\mathrm{diag}\{\eta_1^k, \ldots, \eta_i^k, \ldots, \eta_n^k\}\, sign\Big(g(w^k)\Big), \tag{4}$$

where $sign\Big(g(w^k)\Big)$ denotes the column vector of the signs of the components of $g(w^k)$, $\tau^k > 0$, η_m^k, $m = 1, 2, \ldots, i-1, i+1, \ldots, n$ are small positive real numbers generated by the Rprop learning rates' schedule:

$$if \quad \Big(g_m(w^{k-1}) \cdot g_m(w^k) > 0\Big) \quad then \quad \eta_m^k = min\Big(\eta_m^{k-1} \cdot \eta^+, \Delta_{max}\Big) \tag{5}$$

$$if \quad \Big(g_m(w^{k-1}) \cdot g_m(w^k) < 0\Big) \quad then \quad \eta_m^k = max\Big(\eta_m^{k-1} \cdot \eta^-, \Delta_{min}\Big) \tag{6}$$

$$if \quad \Big(g_m(w^{k-1}) \cdot g_m(w^k) = 0\Big) \quad then \quad \eta_m^k = \eta_m^{k-1}, \tag{7}$$

where $0 < \eta^- < 1 < \eta^+$, Δ_{max} is the learning rate upper bound, Δ_{min} is the learning rate lower bound, and

$$\eta_i^k = -\frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} \eta_j^k \, g_j(w^k) + \delta}{g_i(w^k)}, \qquad 0 < \delta \ll \infty, \qquad g_i(w^k) \neq 0, \tag{8}$$

holds that $\lim_{k\to\infty} \nabla E(w^k) = 0$.

Proof: Evidently, E is bounded below on $\mathbb{R}^n$. The sequence $\{w^k\}_{k=0}^{\infty}$ generated by the iterative scheme (4) follows the direction

$$d^k = -\mathrm{diag}\{\eta_1^k, \ldots, \eta_i^k, \ldots, \eta_n^k\}\, sign\Big(g(w^k)\Big),$$

which is a descent direction if η_m^k, $m = 1, 2, \ldots, i-1, i+1, \ldots, n$ are positive real numbers derived from Relations (5-7), and η_i^k is given by Relation (8), since $g(w^k)^\top d^k < 0$. Following the proof of [12, Theorem 6], since d^k is a descent direction and E is continuously differentiable and bounded below along the radius

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 1-6

$\{w^k + \tau d^k \mid \tau > 0\}$, then there always exist τ^k satisfying (2)–(3) [1, 4]. Moreover, the Wolfe's Theorem [1, 4] suggests that if the cosine of the angle between the descent direction d^k and the $-g(w^k)$ is positive then $\lim_{k\to\infty} g(w^k) = 0$. In our case, indeed $\cos\theta_k = \frac{-g(w^k)^\top d^k}{\|g(w^k)\|\,\|d^k\|} > 0$. $\square$

The modified Rprop, named *GRprop*, is implemented through Relations (4)-(8). The role of δ is to alleviate problems with limited precision that may occur in simulations, and should take a small value proportional to the square root of the relative machine precision. In our tests we set $\delta = 10^{-6}$ in an attempt to test the convergence accuracy of the proposed strategy. Also $\tau^k = 1$ for all k allows the minimisation step along the resultant search direction to be explicitly defined by the values of the local learning rates. The length of the minimisation step can be regulated through τ^k tuning to satisfy (2)–(3). Checking (3) at each iteration requires additional gradient evaluations; thus, in practice (3) can be enforced simply by placing the lower bound on the acceptable values of the learning rates [3, p.1772], i.e. Δ_{min}.

3 Empirical Study

A simple problem is used first to visualise the behavior of the GRprop and compare it with the original method. It is a single node with two weights and logistic activation function. Figure 1 (top row) shows that under the same initial weights and heuristic values, [9], GRprop locates at the center of the contour plot the feasible minimum successfully (Figure 1, left), while Rprop oscillates around the neighbourhood of the minimiser (Figure 1, right). Figure 1 (second row) shows how GRprop locates the minimiser successfully, whilst Rprop's trajectory leads to a point with error value higher than the minimiser.

Below, we report results from 100 independent trials for two problem from the UCI Repository of Machine Learning Databases of the University of California. These 100 random weight initializations are the same for all the learning algorithms, and the training and testing sets were created according to PROBEN1 [8]. The statistical significance of the results has been analysed using the Wilcoxon test [10]. All statements refer to a significance level of 0.05. In all cases we have used networks with sigmoid hidden and output nodes.

The first benchmark is known as the *genes* problem. The data set consists of 1588 patterns. We have used a 120-4-2-3 nodes network as suggested in PROBEN1. Table 1 shows the average performance in terms of: learning speed (Time, secs), convergence success out of the 100 runs (Conv., percentage), and generalisation (Gen., percentage of correctly classified test patterns); a "+" indicates statistical significance of the GRprop results over another method). For example, GRprop-trained networks always generalise slightly better than other networks. Figure 2 (leftside), shows how GRprop converges to a feasible solution ($E < 10^{-5}$), while Rprop to a minimiser with higher error value.

The second task is to decide whether the patient's thyroid has over function, normal function, or under function. We use the *thyroid1* dataset (3600 patterns),

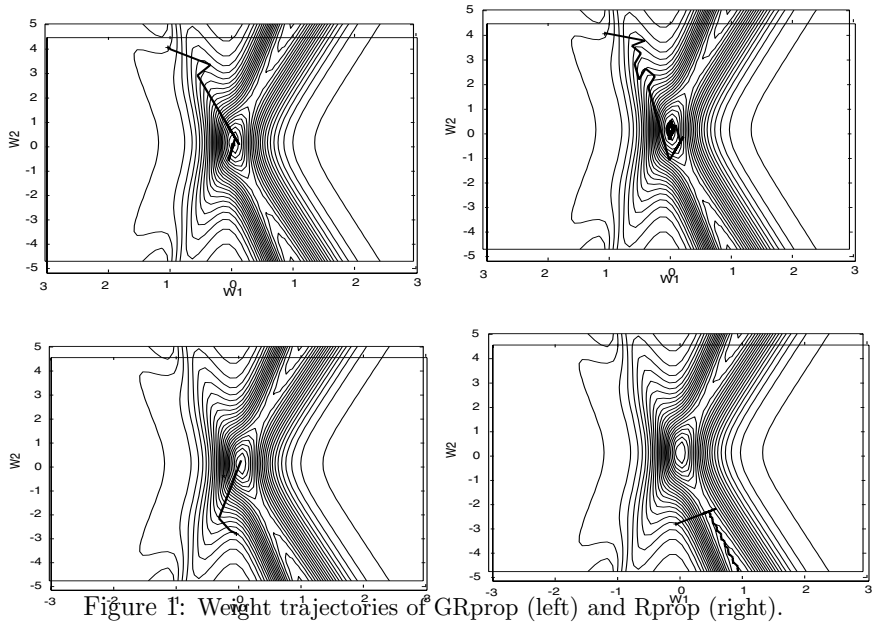

Figure 1: Weight trajectories of GRprop (left) and Rprop (right).

Table 1: Average algorithm performance in the genes and thyroid problems

	Genes			**Thyroid**		
Algorithm	Time	Gen.	Conv.	Time	Gen.	Conv.
Rprop	41.82 (+)	99.1 (+)	97	19.89 (+)	98.12 (+)	87
IRprop	41.04 (+)	99.1 (+)	97	19.58 (+)	98.12 (+)	87
GRprop	36.80	100	100	11.80	98.23	100

and a network with 21-4-3 nodes, as suggested in [11]. Results are given in Table 1. GRprop outperforms the other algorithms particularly in learning speed. Figure 2 (rightside) illustrates a case where GRprop converges to a minimiser with $E < 10^{-5}$ while Rprop gets stuck to a minimiser with higher error value.

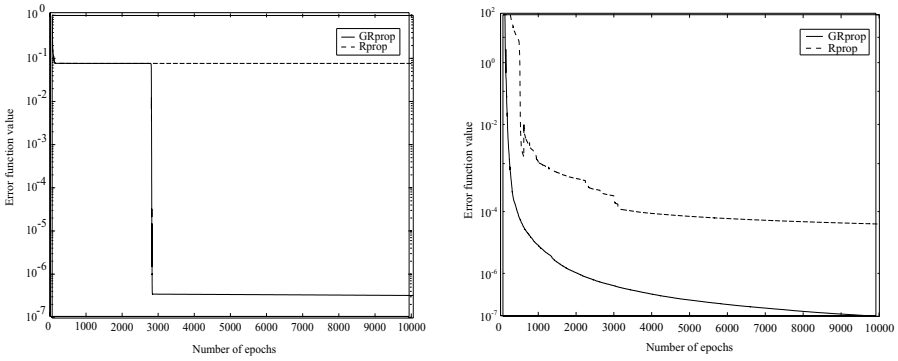

Figure 2: GRprop and Rprop learning curves: genes (left) and thyroid (right).

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 1-6

4 Concluding remarks

In this paper we introduced GRprop, which constitutes and efficient improvement of the original Rprop built on a theoretical basis. We reported comparative results in two benchmark problems. Additional tests have been performed on other UCI benchmarks (e.g. cancer, ecoli and yeast). In our tests GRprop exhibited better convergence speed and stability than Rprop and IRprop.

References

[1] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and nonlinear equations*, SIAM, Philadelphia, 1996.

[2] C. Igel and M. Husken, Empirical evaluation of the improved Rprop learning algorithms, *Neurocomputing*, 50, 105–123, 2003.

[3] G. D. Magoulas, M. N. Vrahatis, and G.S. Androulakis, Improving the Convergence of the Backpropagation Algorithm Using Learning Rate Adaptation Methods, *Neural Computation*, 11, 1999, 1769–1796.

[4] J. Nocedal, Theory of algorithms for unconstrained optimization, *Acta Numerica*, 199–242, 1992.

[5] L. M. Patnaik and K. Rajan, Target detection through image processing and resilient propagation algorithms, *Neurocomputing*, 35, 1-4, 2000, 123–135.

[6] M. Pfister and R. Rojas, Speeding–up backpropagation – A comparison of orthogonal techniques, *Proc. of Joint Conf. Neural Networks*, Nagoya, 517–523, 1993.

[7] M. Pfister and R. Rojas, Qrprop-a hybrid learning algorithm which adaptively includes second order information, *Proc. 4th Dortmund Fuzzy Days*, 55–62, 1994.

[8] Prechelt, L, , PROBEN1-A set of benchmarks and benchmarking rules for neural network training algorithms, Technical report 21/94, Fakultat fur Informatik, Universitat Karlsruhe, 1994.

[9] M. Riedmiller and H. Braun, A direct adaptive method for faster backpropagation learning: The RPROP algorithm, *Proc. Int. Conf. Neural Networks*, San Francisco, 586–591, 1993.

[10] G. Snedecor and W. Cochran, Statistical Methods, Iowa State University Press, 8th edition, 1989.

[11] N. K. Treadgold and T. D. Gedeon, Simulated Annealing and Weight Decay in Adaptive Learning: The SARPROP Algorithm, *IEEE Trans. Neural Networks*, 9, 4, 662–668, 1998.

[12] M.N. Vrahatis, G.D. Magoulas and V.P. Plagianakos, From linear to nonlinear iterative methods, *Applied Numerical Mathematics*, 45, 59-77, 2003.

[13] P. Wolfe, Convergence conditions for ascent methods, *SIAM Review*, 11, 226–235, 1969.

[14] P. Wolfe, Convergence conditions for ascent methods. II: Some corrections, *SIAM Review*, 13, 185–188, 1971.