Training Neural Networks using Two-Point Stepsize Gradient Methods

D.G. Sotiropoulos*, **A.E. Kostopoulos,** and **T.N. Grapsa**

University of Patras, Department of Mathematics, GR–265 04 Rio, Patras, Greece

1 Introduction

Mathematically, the standard "training" problem reduces to finding a set of weights, w, to minimize the error function $E(w)$, defined as the sum of the squares of the errors in the outputs. The weight update equation for any training algorithm has the iterative form:

$$w_{k+1} = w_k + \eta_k \, d_k, \qquad k = 0, 1, 2, \ldots \tag{1}$$

where $w_0 \in \mathbb{R}^n$ is a given starting point, η_k is a stepsize (or learning rate) with $\eta_k \geq 0$, and d_k is a search direction which satisfies $g_k^T \, d_k \leq 0$. The gradient $g_k = \nabla E(w_k)$ easily can be obtained by means of back propagation of errors through the layers. Many local minimization methods have been applied to training feedforward neural networks. Examples include back-propagation (BP), conjugate-gradient and quasi-Newton's methods. BP method minimizes the error function using the steepest descent, namely $d_k = -\nabla E(w_k)$, with fixed (heuristically chosen) stepsize η. While useful for training networks, in practice, BP algorithm can exhibit oscillatory behavior, even with a small stepsize, when it encounters steep valleys. Moreover, convergence is often to a local minimum rather than a global one, since the terrain modelled by the error function in its weight space can be extremely rugged and has many local minima. Something that is inherent in the problem and not restricted to BP.

In this work we develop a sophisticated back-propagation method that automatically adapt the stepsize utilizing information from two points. The search direction is always the negative gradient direction, but for the choice of the stepsize eigenvalues estimates are utilized. Moreover, we derive a two-point stepsize by approximating the new secant equation of Zhang et al. [4] which uses both gradients and function values. In the sequel, we study sufficiently conditions to predict strict descent for both $\nabla E(w)$ and $E(w)$. A switching mechanism is incorporated into the algorithm so that the appropriate stepsize to be chosen. We discuss the properties of the algorithm and establish global convergence under mild assumptions.

2 The Barzilai and Borwein gradient method

Barzilai and Borwein [1] describe a steepest descent method

$$w_{k+1} = w_k - \eta_k \, g_k, \qquad k = 0, 1, 2, \ldots \tag{2}$$

where g_k is the gradient vector of E at w_k and the scalar η_k is given by

$$\eta_k = s_{k-1}^T s_{k-1} / s_{k-1}^T y_{k-1} \tag{3}$$

* Corresponding author: E-mail: dgs@math.upatras.gr, Phone: +30 2610 997 332, Fax: +30 2610 992 965

where $s_{k-1} = w_k - w_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$. This choice of stepsize is proved to be very efficient and numerical experiments indicate that convergence is much more rapid if compared with the optimum steepest descent method that performs exact one-dimensional minimization. However, $E(w_k)$ is not monotonically decreasing, so there are doubts about reliability.

Raydan [3] suggests that a nonmonotone line search technique fits nicely with the nonmonotone behavior of the Barzilai and Borwein and by this way global corvergence can be established. The nonmonotone Armijo condition is given by

$$E\left(w_k + \eta_k \nabla E(w_k)\right) \leq \max_{0 \leq j \leq \min\{k, M-1\}} E(w_{k-j}) - \sigma\, \eta_k \|\nabla E(w_k)\|^2 \tag{4}$$

where the integer M controls the amount of monotonicity that is allowed and $\sigma \in (0,1)$ is a small positive number. This condition allows the acceptance of any point that sufficiently improves the largest of the most recent function values.

3 Derivation of new stepsizes

Quasi-Newton methods approximate the Hessian $\nabla^2 E(w_{k-1})$ by an approximation matrix B_{k-1} so that the new matrix B_k satisfies the secant equation $B_k\, s_{k-1} = y_{k-1}$. Zhang et al. [4] derived a new secant equation which used both available gradient and function value information. The new secant equation is

$$B_k\, s_{k-1} = \tilde{y}_{k-1}, \tag{5}$$

$$\tilde{y}_{k-1} = y_{k-1} + \gamma\, s_{k-1}/\|s_{k-1}\|^2, \tag{6}$$

$$\gamma = 3(g_k + g_{k-1})^T s_{k-1} - 6(E_k - E_{k-1}). \tag{7}$$

In order to determine the new stepsizes, we use the secant equation (5) and we assume that $B_k = \lambda_k I$, where $\lambda_k \in \mathbb{R}$ minimizes $\|s_{k-1} - B_k^{-1}\, \tilde{y}_{k-1}\|$ or $\|B_k\, s_{k-1} - \tilde{y}_{k-1}\|$ with respect to λ_k. Hence, we derive the stepsizes

$$\lambda_k = (\gamma + s_{k-1}^T y_{k-1})/s_{k-1}^T s_{k-1}, \quad k = 2, 3, 4, \ldots \tag{8}$$

or

$$\lambda_k = s_{k-1}^T s_{k-1}/(\gamma + s_{k-1}^T y_{k-1}) \quad k = 2, 3, 4, \ldots \tag{9}$$

respectively, for use in the formula (2).

Remark. If the error function $E(w)$ is quadratic on the line segment between w_{k-1} and w_k, then we have $\eta_k = \lambda_k$. Additionally, if the error function $E(w)$ is strictly convex, then we have that $0 \leq \lambda_k \leq 2\, \eta_k$.

A detailed examination of the numerical progress of the Barzilai and Borwein method (2) and (3) supports the view that the η_k are Rayleigh quotient estimates of some value $1/\lambda_i$. In fact, keeping in mind the standard secant equation, $B_k s_{k-1} = y_{k-1}$, we can verify that the Rayleigh quotient

$$R_q = s_{k-1}^T \left[\int_0^1 \nabla^2 E(w_{k-1} + \tau s_{k-1})\, d\tau \right] s_{k-1} / s_{k-1}^T s_{k-1}$$

corresponds to the average of the Hessian matrix on the line segment between w_{k-1} and w_k. Therefore, R_q is a good approximation to the eigenvalue λ_i for which s_{k-1} is the corresponding eigenvector. So, it is evident that the stepsize (3) is related to the eigenvalues of the Hessian at the minimizer and not to the error function value.

The proposed method composes stepsizes (3) and (8). The choice of the appropriate stepsize at a particular point is our main concern. To accomplish this a switching mechanism is incorporated into the algorithm so that the appropriate stepsize to be chosen according to the status of the current iterative point. To this end, we establish conditions based on available eigenvalue estimates, given from (3), to predict strict descent in both $\nabla E(w)$ and $E(w)$.

4 Model training algorithm

At this point, we briefly give the main steps of the proposed training algorithm.
ALGORITHM 1.

1. Initialize by setting the number of epochs $k = 0$, the weight vector w^0, the stepsize η_0, the error goal eg, the $\sigma \in (0, 1)$ and $M \geq 1$.
2. Compute the weight vector w_1 according to relation (2) and set $k = k + 1$.
3. Compute the new stepsize according to relation (3) or (8).
4. If the stepsize acceptability condition (4) is fulfilled go to Step 6.
5. Set $\eta_k = \eta_k/2$ and go to Step 3.
6. Check if $E(w_{k+1}) > eg$, set $k = k+1$ and go to Step 2. Otherwise, get the final weight vector w_* and the corresponding error function value $E(w_*)$.

The convergence properties of Algorithm 1 are stated in the following theorem.

Theorem 4.1 *Assume that the level set $\mathcal{L}_0 = \{w \in \mathbb{R}^n : E(w) \leq E(w_0)\}$ is bounded and $\nabla E(w)$ is Lipschitz continuous in some neighborhood $\mathcal{N}$ of $\mathcal{L}_0$. Let $\{w_k\}_{k=0}^{\infty}$ be the sequence generated by the Algorithm 1. Then either $\nabla E(w_k) = 0$ for some finite k, or the following properties hold:*
1. $\lim_{k\to\infty} \|\nabla E(w_k)\| = 0$,
2. no limit point of $\{w_k\}$ is a local maximum of E,
3. if the number of stationary points of E in $\mathcal{L}_0$ is finite, then the sequence $\{w_k\}$ converges.

5 Numerical Experiments

The proposed algorithm Adaptive Barzilai-Borwein (ABB) is compared with the original Barzilai-Borwein (BB) as well as with the classical gradient methods from Matlab Neural Network Toolbox.

Table 1 Comparative Results for the XOR Problem

Algorithm	min	max	mean	s.d.	succ.
BP	31	975	100.9	124.8	71.1%
BPM	24	962	128.3	161.8	75.7%
ABP	19	865	47.3	69.1	70.8%
BB	7 4	998 595	65.3 40.6	131.4 79.4	82.7%
ABB	7 6	931 556	56.1 39.2	125.1 77.7	82.7%

Table 2 Comparative Results for the 3-bit Parity Problem

Algorithm	min	max	mean	s.d.	succ.
BP	-	-	-	-	0.0%
BPM	178	995	486.1	203.3	52.6%
ABP	426	959	602.9	123.4	48.5%
BB	50	999	233.5	192.3	72.9%
	33	615	147.1	115.3	
ABB	50	983	222.4	185.5	76.0%
	33	619	148.4	114.7	

The basic steps of (ABB) and (BB) algorithms are identical, except that a switching mechanism is applied in Step 3 in determining the appropriate stepsize to be used at a particular point. The training algorithms have been implemented in Matlab version 6.5. Preliminary results for the XOR and the 3-bit Parity Problems are summarized in Tables 1 and 2, respectively.

References

[1] J. Barzilai and J.M. Borwein. Two point step size gradient methods. *IMA J. Numer. Anal.*,8:141–148, 1988.

[2] L. Grippo, F. Lampariello and S. Lucidi. A nonmonotone line search technique for Newton's method. *SIAM J. Numer. Anal.*,23:707–716, 1986.

[3] M. Raydan. The Barzilai and Borwein gradient method for the large scal unconstrained minimization problem. *SIAM J. Optim.*,7:26–33, 1997.

[4] J.Z. Zhang, N.Y. Deng and L.H. Chen. New quasi–Newton equation and related methods for unconstrained optimization. *Journal of Optimization Theory and Applications*,102:147–167, 1999.