The Third AFSS(1998), 107-112

BACKPROPAGATION BASED ON THE CONJUGATE GRADIENT METHOD WITH THE LINEAR SEARCH BY ORDER STATISTICS AND GOLDEN SECTION

Sang Woong CHOE [a], Jin Choon LEE [b]

a. Dept. of Management Information System, Pohang Junior College
Hunghae-Up, Pohang, Kyungpook, 791-940, Korea
Tel: +82-562-45-1198

b. Dept. of Industrial Engineering, Kyungil University
Hayang, Kyungsan, Kyungpook, 712-701, Korea
Tel: +82-53-850-7266, Fax: +82-53-850-7275, E-mail: jinclee@bear.kyungil.ac.kr

Abstract

In this paper, we propose a new paradigm(NEW_BP) to be capable of overcoming limitations of the traditional backpropagation(OLD_BP). NEW_BP is based on the method of conjugate gradients with the normalized direction vectors and computes step size through the linear search which may be characterized by order statistics and golden section. Simulation results showed that NEW_BP was definitely superior to both the stochastic OLD_BP and the deterministic OLD_BP in terms of accuracy and rate of convergence and might surmount the problem of local minima. Furthermore, they confirmed us that the stagnant phenomenon of training in OLD_BP resulted from the limitations of its algorithm in itself and that unessential approaches would never cured it of this phenomenon.

Keywords : conjugate gradient, normalized direction vector, order statistics, golden section

1. Introduction

The publication of backpropagation method in the late 1980's made a large contribution to the field of neural networks. In retrospect, the great mass of researchers have tended to blindly accept this method for many subsequent years.

The traditional backpropagation method(we call this OLD_BP as shown under) implys the steepest descent method with no linear search(line search, one-dimensional search, step sizing). The direction of steepest descent method is always along the negative gradient. In other words, the negative gradient gives the direction of the most rapid decrease of function value.

But you remember that the new gradient is orthogonal to the direction just traversed. Then the steepest descent method approaches the minimum in a zigzag fashion, which does not, in general, take you to the minimum.

Futhermore, OLD_BP takes a step size(learning-rate) to be some fixed value which is reasonably a small number - on the order of 0.05 to 0.25 - to ensure that the function will converge to a solution. As can be imagined, there are a lot of a fixed step size, i.e., with no linear search. A small value of step size requires a large number of iterations. And the network may bounce around too far from the actual minimum if too large.

Convergence of OLD_BP may be improved by the use of momentum that is a valuable modification to the original method. This additional term tends to keep changes in the same direction. However, a zigzag manner still occurs and it is difficult to choose the fraction of the previous direction added to the current gradient.(momentum-rate)

The most serious problem in OLD_BP training is the existence of local minima, where the error at the network outputs may still be unacceptably high. Local minima problem may be fixed by the judicious use of white noise.(stochastic term) However, it is not a perfect solution to this problem. Also, determining the initial temperature is not easy.

It is undeniable that most researchers have been inclined to manipulate the number of hidden nodes, step size, momentum-rate, initial temperature and initial weights in order to fix local minima problem. To our regret, these manipulations may be unessential on the ground that there is no guarantee that the network will works well.

Consequently, you have seen that the steepest descent method with no linear search used in OLD_BP is not good learning algorithm, even with momentum and white noise.

Now, we propose a new paradigm(the proposed backpropagation) to be capable of overcoming limitations of OLD_BP. The proposed backpropagation(we call this NEW_BP as shown under) is based on the conjugate gradient method with normalized direction vectors and performs the linear search characterized by order statistics and golden section.

The method of conjugate gradients is a special case of the method of conjugate directions and may be defined by the method of proceeding not down the new gradient, rather than in a direction that is somehow constructed to be conjugate to the old gradient and to all previous directions traversed.

In connection with linear search, the closed form of approximate to step size is first derived by the use of order statistics and then once

a iteration, two arguments included in this closed form are computed with the method of golden section.

2. Notation

To describe NEW_BP algorithm, we introduce the following notation.

[1] r -- * layer : input layer(0), hidden layer($1\sim(L-1)$), and outputlayer(L).

[2] p, k -- * the p_th training data, the k_th iteration. : $1 \le p \le P,\ k \ge 0$

[3] N_r -- * number of the $r-$layer nodes except bias term. n_r -- * the n_r_th $r-$layer node. : $0 \le n_r \le N_r\ (r \ne L)$ and bias term, if $n_r = N_r$.

[4] x_{pn_0} -- * value of the p_th training data to the n_0_th input layer node. y_{pn_L} -- * value of the p_th training data to the n_L_th output layer node.

[5] $w^r_{n_r n_{r-1}},\ d^r_{n_r n_{r-1}}$ -- * weight and direction on the connection from the n_{r-1}_th $(r-1)-$layer node to the n_r_th $r-$layer node. : $n_r \ne N_r$

[6] $net^r_{pn_r},\ f^r_{n_r}(net^r_{pn_r})$ -- * net value and output value of the n_r_th $r-$layer node for the p_th training data.

[7] $f^0_{n_0}(net^0_{pn_0}) = net^0_{pn_0} = x_{pn_0}$ -- * $n_0 \ne N_0$

$f^r_{n_r}(net^r_{pn_r}) = f^r_{n_r}(0) = 1$ -- * $n_r = N_r$

$$f^r_{n_r}(net^r_{pn_r}) = f^r_{n_r}\Big(\sum_{n_{r-1}} w^r_{n_r n_{r-1}} f^{r-1}_{n_{r-1}}(net^{r-1}_{pn_{r-1}})\Big) = f^r_{n_r}\Big(\sum_{n_{r-1}\ne N_{r-1}} (w^r_{n_r n_{r-1}} f^{r-1}_{n_{r-1}}(net^{r-1}_{pn_{r-1}})) + w^r_{n_r N_{r-1}}\Big)$$

-- * $r \ne 0,\ n_r \ne N_r$

[8] $Vec(A=[a_{ij}]_{I\times J}) = [b_h]_{IJ\times 1}$ -- * $h = jI + i$, where j is a maximum integer less than or equal to $\frac{h}{I}$ and $0 \le h \le (IJ-1),\ 0 \le i \le (I-1),\ 0 \le j \le (J-1)$.

[9] $E(\theta) = E(Vec(W^1), Vec(W^2), \ldots, Vec(W^L))$

$$= \frac{1}{2}\sum_p\sum_{n_L}[y_{pn_L} - f^L_{n_L}(\sum_{n_{L-1}} w^L_{n_L n_{L-1}} f^{L-1}_{n_{L-1}}(\sum_{n_{L-2}} w^{L-1}_{n_{L-1} n_{L-2}} f^{L-2}_{n_{L-2}}(\ldots\ldots w^3_{n_3 n_2} f^2_{n_2}(\sum_{n_1} w^2_{n_2 n_1} f^1_{n_1}(\sum_{n_0} w^1_{n_1 n_0} f^0_{n_0}(net^0_{pn_0})))\ldots\ldots))))]^2$$

-- * objective(error) function of the network. : $w^r_{N_r n_{r-1}} = 0\ (r \ne 0, L)$

[10] $O^r = [0]_{N_r\times(N_{r-1}+1)},\quad W^r = [w^r_{n_r n_{r-1}}]_{N_r\times(N_{r-1}+1)},$

$D^r = [d^r_{n_r n_{r-1}}]_{N_r\times(N_{r-1}+1)},\quad G^r = [g^r_{n_r n_{r-1}}]_{N_r\times(N_{r-1}+1)}.$

-- * $r-$layer zero, weight, direction and gradient matrix respectively. : $r \ne 0,\ n_r \ne N_r$

[11] $\theta^r = [Vec(O^1), \ldots, Vec(W^r), Vec(O^{r+1}), \ldots, Vec(O^L)]^T$

$\lambda^r = [Vec(O^1), \ldots, Vec(D^r), Vec(O^{r+1}), \ldots, Vec(O^L)]^T$

$\rho^r = [Vec(O^1), \ldots, Vec(G^r), Vec(O^{r+1}), \ldots, Vec(O^L)]^T$

$\sum_r \theta^r = \theta,\ \sum_r \lambda^r = \lambda,\ \sum_r \rho^r = \rho,\ r \ne 0$

[12] $\langle\theta^r\rangle^T = [Vec^T(O^1), \ldots, Vec^T(W^r), Vec^T(O^{r+1}), \ldots, Vec^T(O^L)]$

$\langle\lambda^r\rangle^T = [Vec^T(O^1), \ldots, Vec^T(D^r), Vec^T(O^{r+1}), \ldots, Vec^T(O^L)]$

$\langle\rho^r\rangle^T = [Vec^T(O^1), \ldots, Vec^T(G^r), Vec^T(O^{r+1}), \ldots, Vec^T(O^L)]$

$\sum_r\langle\theta^r\rangle^T = \theta^T,\ \sum_r\langle\lambda^r\rangle^T = \lambda^T,\ \sum_r\langle\rho^r\rangle^T = \rho^T,\ r \ne 0$

[13] $g^r_{n_r n_{r-1}} = \dfrac{\partial E(\cdot)}{\partial w^r_{n_r n_{r-1}}} = -\sum_p \delta^r_{pn_r} f^{r-1}_{n_{r-1}}(net^{r-1}_{pn_{r-1}})$

$$\delta^r_{pn_r} = \frac{df^r_{n_r}(net^r_{pn_r})}{dnet^r_{pn_r}}\sum_{n_{r+1}}\delta^{r+1}_{pn_{r+1}} w^{r+1}_{n_{r+1} n_r},\ r \ne 0, L \text{ and}$$

$$\delta^L_{pn_L} = \frac{df^L_{n_L}(net^L_{pn_L})}{dnet^L_{pn_L}}(y_{pn_L} - f^L_{n_L}(net^L_{pn_L}))$$

[14] $\nabla E(\theta) = \rho = \left[\dfrac{\partial E(\theta)}{\partial Vec(W^r)}\right]_{L\times 1},\quad \dfrac{\partial E(\theta)}{\partial Vec(W^r)} = Vec(G^r)$

-- * gradient vector of $E(\theta)$.

[15] $H(\theta) = [H_{ab}]_{L\times L},\ H_{ab} = \dfrac{\partial^2 E(\theta)}{\partial Vec(W^b)\,\partial Vec(W^a)} = \left[\dfrac{\partial^2 E(\theta)}{\partial w^b_{n_b n_{b-1}}\partial w^a_{n_a n_{a-1}}}\right]_{p\times q}$

-- * positive definite Hessian matrix. : $p = N_a\times(N_{a-1}+1),\ q = N_b\times(N_{b-1}+1)$

[16] $\|x\| = (x^T x)^{1/2}$

3. NEW_BP : Algorithm and step size

3.1 Algorithm

Normalized direction vectors, λ^r_k, are generated by

$$\lambda^r_k = \frac{(-\rho^r_k + \eta^r_{k-1}\lambda^r_{k-1})}{\|-\rho^r_k + \eta^r_{k-1}\lambda^r_{k-1}\|},\ k \ge 1$$
$$\lambda^r_0 = \frac{(-\rho^r_0)}{\|-\rho^r_0\|} \qquad (1)$$

where η^r_{k-1} is a scalar and $r \ne 0$. Therefore,

$$Vec(D^r_k) = \frac{(-Vec(G^r_k) + \eta^r_{k-1}Vec(D^r_{k-1}))}{\|-Vec(G^r_k) + \eta^r_{k-1}Vec(D^r_{k-1})\|},\ k \ge 1$$
$$Vec(D^r_0) = \frac{(-Vec(G^r_0))}{\|-Vec(G^r_0)\|} \qquad (2)$$

A scalar η^r_{k-1} is chosen to make λ^r_k $H(\theta_k)-$conjugate to λ^r_{k-1} ; that is, it is given as

$$\eta^r_{k-1} = \frac{\langle\rho^r_k\rangle^T H(\theta_{K-1})\lambda^r_{k-1}}{\langle\lambda^r_{k-1}\rangle^T H(\theta_{K-1})\lambda^r_{k-1}} = \frac{Vec^T(G^r_k)H_n(\theta_{K-1})Vec(D^r_{k-1})}{Vec^T(D^r_{k-1})H_n(\theta_{K-1})Vec(D^r_{k-1})},\ k \ge 1 \qquad (3)$$

The next point θ^r_{k+1} is generated from θ^r_k by the linear search

in the direction of λ_k^r and is expressed as

$$\theta_{k+1}^r = \theta_k^r + \tau_k^r \lambda_k^r , \; k \geq 0 \tag{4}$$

Therefore,

$$\mathrm{Vec}(W_{k+1}^r) = \mathrm{Vec}(W_k^r) + \tau_k^r \mathrm{Vec}(D_k^r) , \; k \geq 0 \tag{5}$$

where τ_k^r is a step size and we select it such that

$$\begin{aligned} \mathrm{Min}_{\tau_k} \{ & E(\theta_k^{+r} = \sum_{i \neq r}^{L} \theta_k^i + \theta_{k+1}^r) = \\ & E(\mathrm{Vec}(W_k^1), \ldots, \mathrm{Vec}(W_{k+1}^r), \ldots, \mathrm{Vec}(W_k^L))\} \end{aligned} \tag{6}$$

Let $\theta_k \in S$ which is an open subset in R^n ; $E(\theta_k^{+r})$ can be expanded into a Taylor series about θ_k

$$\begin{aligned} & \mathrm{Min}_{\tau_k} E(\theta_k^{+r}) \\ & \quad \cong \mathrm{Min}_{\tau_k} \left\{ E(\theta_k) + \langle \rho_k^r \rangle^T \lambda_k^r \tau_k^r + \frac{1}{2} \langle \lambda_k^r \rangle^T H(\theta_k) \lambda_k^r (\tau_k^r)^2 \right\} \end{aligned} \tag{7}$$

From Eq. (7),

$$\tau_k^r = \frac{-\langle \rho_k^r \rangle^T \lambda_k^r}{\langle \lambda_k^r \rangle^T H(\theta_K) \lambda_k^r} = \frac{-\mathrm{Vec}^T(G_k^r)\mathrm{Vec}(D_k^r)}{\mathrm{Vec}^T(D_k^r) H_m(\theta_K) \mathrm{Vec}(D_k^r)} \tag{8}$$

Now, we can see

$$\frac{1}{\tau_{k-1}^r} (\rho_k^r - \rho_{k-1}^r) \cong H(\theta_{k-1}) \lambda_{k-1}^r \tag{9}$$

By substituting Eq. (9) into Eq. (3), we have an alternate form for η_{k-1}^r

$$\begin{aligned} \eta_{k-1}^r & \cong \frac{\langle \rho_k^r \rangle^T (\rho_k^r - \rho_{k-1}^r)}{\langle \lambda_{k-1}^r \rangle^T (\rho_k^r - \rho_{k-1}^r)} \\ & = \frac{\mathrm{Vec}^T(G_k^r)(\mathrm{Vec}(G_k^r) - \mathrm{Vec}(G_{k-1}^r))}{\mathrm{Vec}^T(D_{k-1}^r)(\mathrm{Vec}(G_k^r) - \mathrm{Vec}(G_{k-1}^r))} \end{aligned} \tag{10}$$

Also, from Eq. (8) and (9), we have Eq. (11).

$$\langle \rho_k^r \rangle^T \lambda_{k-1}^r \cong 0 \tag{11}$$

From Eq. (11),

$$\langle \rho_k^r \rangle^T \lambda_k^r \cong \frac{(-\langle \rho_k^r \rangle^T \rho_k^r)}{\| -\rho_k^r + \eta_{k-1}^r \lambda_{k-1}^r \|} \tag{12}$$

From Eq. (12),

$$\begin{aligned} & \tau_{k-1}^r \langle \lambda_{k-1}^r \rangle^T H(\theta_{k-1}) \rho_{k-1}^r \\ & \cong \eta_{k-2}^r \tau_{k-1}^r \langle \lambda_{k-1}^r \rangle^T H(\theta_{k-1}) \lambda_{k-2}^r - \langle \rho_{k-1}^r \rangle^T \rho_{k-1}^r \end{aligned} \tag{13}$$

Consequently, from Eq. (9) and (13), we obtain

$$\begin{aligned} \langle \rho_k^r \rangle^T \rho_{k-1}^r & \cong \eta_{k-2}^r \tau_{k-1}^r \langle \lambda_{k-1}^r \rangle^T H(\theta_{k-1}) \lambda_{k-2}^r \\ & \cong \eta_{k-2}^r \tau_{k-1}^r \langle \lambda_{k-1}^r \rangle^T H(\theta_{k-2}) \lambda_{k-2}^r = 0 \end{aligned} \tag{14}$$

And we can get two alternate expression for η_{k-1}^r ; that is, by substituting Eq. (11) into (10)

$$\begin{aligned} \eta_{k-1}^r & \cong \frac{\langle \rho_k^r \rangle^T (\rho_{k-1}^r - \rho_k^r)}{\langle \rho_{k-1}^r \rangle^T \lambda_{k-1}^r} \\ & = \frac{\mathrm{Vec}^T(G_k^r)(\mathrm{Vec}(G_{k-1}^r) - \mathrm{Vec}(G_k^r))}{\mathrm{Vec}^T(G_{k-1}^r)\mathrm{Vec}(D_{k-1}^r)} \end{aligned} \tag{15}$$

and by substituting Eq. (14) into (15)

$$\eta_{k-1}^r \cong \frac{-\langle \rho_k^r \rangle^T \rho_k^r}{\langle \rho_{k-1}^r \rangle^T \lambda_{k-1}^r} = \frac{-\mathrm{Vec}^T(G_k^r)\mathrm{Vec}(G_k^r)}{\mathrm{Vec}^T(G_{k-1}^r)\mathrm{Vec}(D_{k-1}^r)} \tag{16}$$

As mentioned above, we have obtained three types for η_{k-1}^r. We will adopt type II as the canonical form for η_{k-1}^r of NEW_BP. (see Table 1)

Table 1 Three types for η_{k-1}^r

Type	Eq.	$\langle \rho_k^r \rangle^T \lambda_{k-1}^r \cong 0$	$\langle \rho_k^r \rangle^T \rho_{k-1}^r \cong 0$	Relation
I	(10)	not accepted	not accepted	Hestenes & Stiefel ver.
II	(15)	accepted	not accepted	Polak & Ribiere ver.
III	(16)	accepted	accepted	Fletcher & Reeves ver.

3.2 Step size

3.2.1 Closed form by the use of order statistics

From Eq. (9), Eq. (8) can be written as

$$\tau_k^r \cong \frac{2(E(\theta_k^{+r}) - E(\theta_k))}{\langle \rho_{k+1}^r \rangle^T \lambda_k^r + \langle \rho_k^r \rangle^T \lambda_k^r} \tag{17}$$

since $E(\theta_k^{+r}) \cong E(\theta_k) + \frac{1}{2}(\langle \rho_{k+1}^r \rangle^T \lambda_k^r + \langle \rho_k^r \rangle^T \lambda_k^r)\tau_k^r$.

However, Eq. (17) may not be a directly available expression since we must estimate ρ_{k+1}^r and $E(\theta_k^{+r})$. For this reason, we will grant three desirable properties to $m(\widehat{\tau_k^r})$ which is the mean of $\widehat{\tau_k^r}$, an estimate of τ_k^r with the purpose of a successful training. We can analytically derive the unique $m(\widehat{\tau_k^r})$ which satisfies three properties and then use this as the good approximate to τ_k^r. These three desirable properties are as follows.

[Property 1] θ_{k+1}^r lies in a given direction λ_k^r from θ_k^r ; that is, $Vec(W_{k+1}^r)$ lies in a given direction $Vec(D_k^r)$ from $Vec(W_k^r)$.

[Property 2] $\tau_k^1, \tau_k^2, \cdots, \tau_k^{L-1}, \tau_k^L$ are independent and identically distributed continuous uniform (L_k, U_k) random variables.

[Property 3] $\widehat{\tau_k^1} < \widehat{\tau_k^2} < \cdots\cdots < \widehat{\tau_k^{L-1}} < \widehat{\tau_k^L}$

In short, if Eq. (17) satisfies the above-mentioned properties, it can be written as

$$\widehat{\tau_k^r} \cong \frac{2\left(\widetilde{E(\theta_k^{+r})} - E(\theta_k)\right)}{\langle \rho_k^r \rangle^T \lambda_k^r}, \qquad L_k < \widehat{\tau_k^1} < \widehat{\tau_k^2} < \cdots < \widehat{\tau_k^{L-1}} < \widehat{\tau_k^L} < U_k \tag{18}$$

From Eq. (18), we get

$$\tau_k^r \approx m(\widehat{\tau_k^r}) \cong \frac{2\left\{m\left(\widetilde{E(\theta_k^{+r})}\right) - E(\theta_k)\right\}}{\langle \rho_k^r \rangle^T \lambda_k^r}$$
$$L_k < \tau_k^1 \approx m(\widehat{\tau_k^1}) < \tau_k^2 \approx m(\widehat{\tau_k^2}) < \cdots < \tau_k^{L-1} \approx m(\widehat{\tau_k^{L-1}}) < \tau_k^L \approx m(\widehat{\tau_k^L}) < U_k \tag{19}$$

Let $\tau_k^{(r)}$ be the r_{th} smallest value of $(\tau_k^1, \tau_k^2, \cdots, \tau_k^{L-1}, \tau_k^L)$. By this definition, $\tau_k^{(1)} < \tau_k^{(2)} < \cdots\cdots < \tau_k^{(L-1)} < \tau_k^{(L)}$ are the order statistics corresponding to the random variables $(\tau_k^1, \tau_k^2, \cdots, \tau_k^{L-1}, \tau_k^L)$. Therefore, the density and distribution function of $\tau_k^{(r)}$ are given by

$$f_{\tau_k^{(r)}}(\tau) = \frac{L!(U_k - L_k)^{-L}}{(L-r)!(r-1)!}(\tau - L_k)^{r-1}(U_k - \tau)^{L-r}, \quad \text{where } L_k < \tau < U_k$$
$$F_{\tau_k^{(r)}}(\tau^*) = \int_{L_k}^{\tau^*} f_{\tau_k^{(r)}}(\tau)d\tau, \quad \text{where } L_k \le \tau^* \le U_k \tag{20}$$

By [Property 3],

$$m(\widehat{\tau_k^r}) = m(\tau_k^{(r)}) \approx \tau_k^r \tag{21}$$

From Eq. (20), we obtain

$$m(\tau_k^{(r)}) = \left\{ {}_rL C_r \sum_{x=0}^{L-r} \frac{{}_{L-r}C_x(-1)^x}{x+r+1} \right\} U_k + \left\{ {}_rL C_r \sum_{x=0}^{L-r} \frac{{}_{L-r}C_x(-1)^x}{(x+r)(x+r+1)} \right\} L_k \tag{22}$$

Using Eq. (19) and (20),

$${}_rL C_r \sum_{x=0}^{L-r} \frac{{}_{L-r}C_x(-1)^x}{x+r} = 1 \tag{23}$$

And we can see

$$\sum_{x=0}^{m} \frac{{}_mC_x(-1)^x}{x+1} = \frac{1}{m+1} \tag{24}$$

From Eq. (23) and (24),

$$\sum_{x=0}^{L-r} \frac{{}_{L-r}C_x(-1)^x}{x+r+1} = \frac{1}{(L+1)\,{}_LC_r}, \quad 1 \le r \le L \tag{25}$$

Finally, by Eq. (23) and (25), Eq. (22) is expressible as

$$m(\tau_k^{(r)}) = \left(\frac{r}{L+1}\right)U_k + \left(\frac{L-r+1}{L+1}\right)L_k = m(\widehat{\tau_k^r}) \approx \tau_k^r \tag{26}$$

In conclusion, we have obtained Eq. (26), that is, the closed form of a good approximate to the true step size, τ_k^r.

3.2.2 Computation L_k, U_k by the use of golden section

The golden section requires use of the Fibonacci fractions $\frac{3-\sqrt{5}}{2}$ and $\frac{\sqrt{5}-1}{2}$. A key property of this procedure is that only one new point need be evaluated at each iteration due to the fact that one point has already been determined. In other words, L_k and U_k in Eq. (26) can be determined at each iteration, only compared $E(\theta_k)$ with $E(\theta_{k-1})$.

Let (b_k^L, e_k^L) and (b_k^U, e_k^U) be the intervals where L_k and U_k lie respectively.

If $E(\theta_{k-1}) \geq E(\theta_k)$ for all $k \geq 2$,

(1) $(b_k^L, e_k^L) = (p_{1,k-1}^L, e_{k-1}^L)$, $(b_k^U, e_k^U) = (p_{1,k-1}^U, e_{k-1}^U)$

(2) $p_{1,k-1}^L = b_{k-1}^L + \frac{3-\sqrt{5}}{2}(e_{k-1}^L - b_{k-1}^L)$,
$p_{1,k-1}^U = b_{k-1}^U + \frac{3-\sqrt{5}}{2}(e_{k-1}^U - b_{k-1}^U)$

(3) $p_{1,k}^L = p_{2,k-1}^L = b_{k-1}^L + \frac{\sqrt{5}-1}{2}(e_{k-1}^L - b_{k-1}^L)$,
$p_{1,k}^U = p_{2,k-1}^U = b_{k-1}^U + \frac{\sqrt{5}-1}{2}(e_{k-1}^U - b_{k-1}^U)$

(4) $p_{2,k}^L = L_k = b_k^L + \frac{\sqrt{5}-1}{2}(e_k^L - b_k^L)$,
$p_{2,k}^U = U_k = b_k^U + \frac{\sqrt{5}-1}{2}(e_k^U - b_k^U)$

Therefore, $\tau_k^r \approx \left(\frac{r}{L+1}\right)p_{2,k}^U + \left(\frac{L-r+1}{L+1}\right)p_{2,k}^L$ (27)

If $E(\theta_{k-1}) < E(\theta_k)$ for all $k \geq 2$,

(1) $(b_k^L, e_k^L) = (b_{k-1}^L, p_{2,k-1}^L)$, $(b_k^U, e_k^U) = (b_{k-1}^U, p_{2,k-1}^U)$

(2) $p_{2,k-1}^L = b_{k-1}^L + \frac{\sqrt{5}-1}{2}(e_{k-1}^L - b_{k-1}^L)$,
$p_{2,k-1}^U = b_{k-1}^U + \frac{\sqrt{5}-1}{2}(e_{k-1}^U - b_{k-1}^U)$

(3) $p_{2,k}^L = p_{1,k-1}^L = b_{k-1}^L + \frac{3-\sqrt{5}}{2}(e_{k-1}^L - b_{k-1}^L)$,
$p_{2,k}^U = p_{1,k-1}^U = b_{k-1}^U + \frac{3-\sqrt{5}}{2}(e_{k-1}^U - b_{k-1}^U)$

(4) $p_{1,k}^L = L_k = b_k^L + \frac{3-\sqrt{5}}{2}(e_k^L - b_k^L)$,
$p_{1,k}^U = U_k = b_k^U + \frac{3-\sqrt{5}}{2}(e_k^U - b_k^U)$

Therefore, $\tau_k^r \approx \left(\frac{r}{L+1}\right)p_{1,k}^U + \left(\frac{L-r+1}{L+1}\right)p_{1,k}^L$ (28)

For $k=0$,

$$\tau_0^r \approx \left(\frac{r}{L+1}\right)\left(b_0^U + \frac{3-\sqrt{5}}{2}(e_0^U - b_0^U)\right) + \left(\frac{L-r+1}{L+1}\right)\left(b_0^L + \frac{3-\sqrt{5}}{2}(e_0^L - b_0^L)\right) \quad (29)$$

and for $k=1$,

$$\tau_1^r \approx \left(\frac{r}{L+1}\right)\left(b_1^U + \frac{\sqrt{5}-1}{2}(e_1^U - b_1^U)\right) + \left(\frac{L-r+1}{L+1}\right)\left(b_1^L + \frac{\sqrt{5}-1}{2}(e_1^L - b_1^L)\right) \quad (30)$$

where $b_1^U = b_0^U$, $e_1^U = e_0^U$, $b_1^L = b_0^L$, $e_1^L = e_0^L$.

Suppose $\mathrm{Max}_{k,r}\tau_k^r = \mathrm{Max}_k \tau_k^L < C$, where C is a given positive constant, in general, 1 so as to determine b_0^U, e_0^U, b_0^L and e_0^L.

We intend to maximize $(b_0^U < U_0 < e_0^U) \cap (b_0^L < L_0 < e_0^L)$ subject to

$$\frac{U_0}{L+1} + \frac{L \cdot L_0}{L+1} > 0,\ U_0 > L_0,\ \frac{L \cdot U_0}{L+1} + \frac{L_0}{L+1} < C.$$

Thus, if $b_0^L = 0$, we have

$$b_0^U = e_0^L = \frac{C}{2},\ e_0^U = C\left(1 + \frac{1}{2L}\right) \quad (31)$$

4. Simulation example

(See Table 2)

4.1 First case : XOR problem

(See Tables 3, 4)

4.2 Second case : Evaluation of Rosenbrock function value

(See Tables 3, 5)

5. Concluing remarks

From simulation results, compared with OLD_BP, NEW_BP has three major advantages. These are as follows.

① NEW_BP is by far general, i.e., OLD_BP is a special case of NEW_BP.

② NEW_BP is definitely superior to OLD_BP in terms of accuracy and rate of convergence.

③ NEW_BP may surmount the problem of local minima.

Table 2 Input parameters in four simulations

Simulation	OLD_BP				NEW_BP	
	Step Size	Momentum -Rate	Initial Temperature	Type	C	Type
1	0.3	0.15	0	d*	0.3	II
2	0.2	0.1	0	d	0.2	II
3	0.3	0.15	0.001	s**	0.15	II
4	0.2	0.1	0.001	s	0.1	II

* : deterministic ** : stochastic

Table 3 Simulation environment in two cases

	Input Nodes	First Hidden Nodes	Output Nodes	Output Function	Error Level	Max. Iteration	Initial Weights	Training Data
1_st Case	2	2	1	sigmoid	1E-12	3000	n*	4
2_nd Case	14	10	7	sigmoid	1E-12	5000	n	20

* : normalized

Table 4 Simulation results of 1_st case, XOR problem

Simulation	OLD_BP			NEW_BP		
	Run Time	Iterations	Error	Run Time	Iterations	Error
1	1.43	3000	0.008530726532	0.16	500	0.000000000000
2	1.43	3000	0.095760896064	0.27	800	0.000000000000
3	1.43	3000	0.008250918171	0.38	1000	0.000000000000
4	1.43	3000	0.084207681382	0.49	1400	0.000000000001

Table 5 Simulation results of 2_nd case, Rosenbrock function evaluation

Simulation	OLD_BP			NEW_BP		
	Run Time	Iterations	Error	Run Time	Iterations	Error
1	95	5000	0.007649889714	19.89	1200	0.000000000001
2	95	5000	0.012565657528	29.62	1800	0.000000000000
3	94.84	5000	0.007521271519	37.80	2300	0.000000000001
4	94.84	5000	0.011901285573	57.47	3500	0.000000000000

References

1. S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images", IEEE Proc. Pattern Analysis and Machine Intelligence, PAMI-6, 721-741, 1984.
2. B. Gidas, "Global Optimization via the Langevin Equation", Proc. 24th Conf. on Decision and Control, Ft. Lauderdale, 774-778, 1985.
3. James A. Freeman and David M. Skapura, Neural Networks : Algorithms, Applications and Programming Techniques, Addison Wesley, 1992.
4. M. R. Hestenes and E. Stiefel, "Methods of Conjugate Gradient for Solving Linear Systems", J. Res. Nat. Bur. Standards, 49, 409-436, 1952.
5. R. Fletcher and C. M. Reeves, "Function Minimization by Conjugate Gradients", the Computer Journal 7, 149-153, 1964.
6. E. Polak and G. Ribiere, "Note Surla Convergence de Methodes Conjugees", Revue Francaise Inform. Rech. Operation 16-R1, 35-43.