Convergence of the Improve Levenberg-Marquardt Method For solving Singular Perturbation Problems

Khalid. Mindeel M. Al-Abrahemee[(1)]
Luma. N. M. Tawfiq[(2)]
Department of Mathematics, College of Education / University of AL-Qadisiyhah

Abstract :

In this paper , we overcome and over passing the disadvantages of the standard LM training algorithm for solving Singular Perturbation Problems in the case of whether the matrix $J(w)$rectangular or singular, by the new technique that we suggest SVD of J and J^{-1} . Secondly suggest new calculation of combination coefficient (μ_k) that is, $\mu_k = \|E(w)\|^2$. Consider the nonlinear performance equations E(w) = 0 where E(w): $R^n \to R^m$ is continuously differentiable and has nonempty solution W^* and we refer $\|\cdot\|$ to the 2-norm in all cases .Starting with the suitable choice of the parameter μ, we prove that, if $\|E(w)\|$ gives the error bound for some $w^* \in W$, then the sequence $\{w_k\}$ generated using the modified LM algorithm converges super linearly and quadratically to the solution of equation E(w) = 0.

1. Introduction

All possible solution , can be applied to get the optimal value of weights for feed forward neural networks. Naturally, local searches are generally gave the local solutions; once attempt to avoid this limitation. The performance for training varies depending on the network configuration and error surface for a given problem. Since the gradient information of error surface is available for the most widely applied network configurations, the most popular optimization methods have been variants of gradient based BP algorithms [1]. Of course, this is sometimes the result of an inseparable combination of network configuration and training algorithm which limits the freedom to choose the optimization method. For small networks the problem of local minima can be efficiently avoided by using repeated trainings, changing the length of step size and randomly initialized weight values

[4]. Nevertheless, global optimization algorithms may be useful for validation of an optimal solution achieved by BP training algorithm.

In the training process the parameters of the network are updated as the following rule [1]:

$$w_{k+1} = w_k + \eta \Delta w_k \quad (1)$$

Where,

η: is learning rate;

Δw_k: computed by one of training algorithm.

The learning rate η is used to determine the size of the weight update in most of the training algorithms.

There are several back propagation training algorithms: gradient descent (GD), and gradient descent with momentum. These two methods are often too slow for practical problems [3]. This paper consists some training algorithms and its modification which are high performance process, it can assist to converge faster than the previous back propagation training algorithms.

2. Performance Functions

Performance functions are used in supervised training type to help update the network parameters. Supervised training, ANN is provided with the desired output for each input. The error is defined as the difference between the desired output and the network output (called actual output). Network parameters are updated according to one of two performance functions to reduce the network error [1]:

1. **Least mean of squared errors (MSE)**: minimizes the average of the squared errors.
2. **Least sum of squared errors (SSE)**: minimizes the sums of the squared errors, it is calculated by:

$$E(x, w) = \frac{1}{2} \sum_{P=1}^{P} \sum_{m=1}^{M} e_{P,m}^2 \quad (2)$$

Where,

x: is the input vector;

w: is the weight vector;

$e_{P,m}$: is the training error at output m when applying input p and it is defined as:

$$e_{p,m} = d_{p,m} - t_{p,m} \quad (3)$$

Where,

d: is the desired output vector;

t: is the actual output vector; and

g: is the gradient defined as the first – order partial derivative of total error function (2), i.e.,

$$g = \frac{\partial E(x,w)}{\partial w} = \left[\frac{\partial E}{\partial w_1}\frac{\partial E}{\partial w_2} \ldots\ldots \frac{\partial E}{\partial w_N}\right]^T \quad (4)$$

3. Levenberg-Marquardt Algorithm (trainlm)

Levenberg–Marquardt algorithm (LM) is one of the best training algorithm which has 2nd order convergence without having to compute the Hessian matrix. When the performance function has the form of a sum of squares, then the Hessian matrix can be approximated as J^TJ and the gradient can be computed as $g = J^TE(w_k)$, where J is the Jacobian matrix, which contains first derivatives of the network errors with respect to the parameters (weights and biases), and $E(w_k)$, is a vector of network errors. The Levenberg–Marquardt algorithm uses the following approximation to the Hessian matrix [6]:

$$H \approx J^TJ + \mu I \quad (5)$$

Where;

μ: is always positive, called combination coefficient and I: is the identity matrix.

From equation (5), one may notice that the elements on the main diagonal of the approximation Hessian matrix will be greater than zero.

Therefore, uses this approximation (equation (5) in the following Newton update:

$$w_{k+1} = w_k - (J^TJ + \mu I)^{-1}J_k\, E(w_k), \quad (6)$$

when μ = 0, this is just Newton's method, when μ is large, this becomes gradient descent with a small step size [1]. If μ is very big, it can be interpreted as the learning coefficient in the training process (1): $\eta = \frac{1}{\mu}$

The only drawback related to this training algorithm is the exact evaluation of the Hessian matrix (H) which is computationally intensive. The computation of the inverse Hessian (H^{-1}) is even more computationally intensive [2].

4. Improve Levenberg-Marquardt Training Algorithm

In this thesis, we over passing the drawback of LM algorithm, firstly we suggest SVD of J and J^{-1}, if $J(w)\ is$ a rectangular matrix or singular, then we use SVD of $J(w)$. Secondly, we suggest new calculation of μ_k , that is, $\mu_k = \|E(w)\|^2$.

Consider the nonlinear performance equations:

$$E(w) = 0 \qquad (7)$$

where E(w): $R^n \to R^m$ is continuously differentiable and E(x) is Lipschitz continuous, that is, $\|E(w_2) - E(w_1)\| \leq L\|w_2 - w_1\|$, where L is Lipschitz constant. Suppose that the equation (7) has nonempty solution W^* and we refer $\|\cdot\|$ to the 2-norm in all cases. Starting with the suitable choice of the parameter μ, we prove that, if $\|E(w)\|$ gives the error bound for some $w^* \in W$, then the sequence $\{w_k\}$ generated using the modified LM algorithm converges super linearly to the solution of equation (7).

Now we will take some definitions, hypotheses, axioms and theorems to help the proof of convergence for modified LM algorithm depending on the above condition.

Definition 1 [5]

Let *N* be a subset of R^n such that $N \cap W^* \neq \emptyset$. We say that $\|E(w)\|$ provides a local error bound on *N* for system (7), if there exists a positive constant $c > 0$ such that:

$$\|E(w)\| \geq c\,dist(w, W^*)$$

Note, if $J(w^*)$ is nonsingular depending on solution at w^* of equation (7), then w^* represents the unique solution, hence $\|E(w)\|$ gives an error bound for neighborhood of w^*.

5. Convergence of the Improve Levenberg-Marquardt Method

To study the rate of convergence for the modified LM algorithm, we introduce the following hypotheses.

Hypothesis 2

Suppose that $\|E(w)\|$ is continuously differentiable, and its Jacobian $J(w)$ satisfies Lipschitz condition for neighborhood of $w^* \in W^*$, i.e., ∃ a positive constants L and $b_1 less\ than\ 1$ such that**:**

$$\|J(w_2) - J(w_1)\| \leq L\|w_2 - w_1\| \qquad (8)$$

(a) Let $\|E(w)\|$ have an error bound on $N(w^*, b_1)$ to the equation (7), i.e., ∃ a constant $c_1\ which\ is\ greater\ than\ 0$ such that:

$$\|E(w)\| \geq c_1 dist(w, W^*)\,, \qquad \forall w \in N(w^*, b_1) \qquad (9)$$

By Hypothesis 2.2(a), we have:

$$\|E(w_2) - E(w_1) - J(w_1)(w_2 - w_1)\| \leq L_1\|w_2 - w_1\|^2\,, \forall w_1, w_2 \in N(w^*, b_1) \qquad (10)$$

and, ∃ a constant $L_2 > 0$, such that:

$$\|E(w_2) - E(w_1)\| \leq L_2\|w_2 - w_1\|, \qquad \forall w_1, w_2 \in N(w^*, b_1) \qquad (11)$$

Now, to study the convergence of the modified Levenberg-Marquardt algorithm, starting with the update rule of the weight which is computed by:

$$w_{k+1} = w_k + \rho_k$$

where ρ_k is the search direction given as:

$$\rho_k = -(J(w_k)^T J(w_k) + \mu_k I)^{-1} J(w_k)^T E(w_k) \qquad (12)$$

for simplicity, we use the notations: $E_k = E(w_k)$, $J_k = J(w_k)$ from now on.

Hypothesis 3

Suppose that $\mu_k = \|E_k\|^\delta$ for all *k*, where $\delta \in [1, 2]$. Yamashita and Fukushima in [5] showed the quadratic convergence of the Levenberg-Marquardt method for nonsingular system based on the analyses of an unconstrained optimization problem. Here, we prove the super linear convergence of the modified LM algorithm (second type) which by choosing $\mu_k = \|E(w_k)\|^2$, then, obtain the convergence of the algorithm depending on the SVD for the Jacobian matrix.We denote $\bar{w}_k$ the vector in W^* that satisfies: $\|w_k - \bar{w}_k\| = dist(w_k, W^*)$

Lemma 5.4

If the hypotheses 2 and 3 are satisfied and $w_k \in N\left(w^*, \frac{b_1}{2}\right)$, then ∃ a constant $c_2 > 0$ such that: $\|\rho_k\| \leq c_2 dist(w_k, W^*)$ (13)

Proof

Since $w_k \in N\left(w^*, \frac{b_1}{2}\right)$, we have:

$$\|\bar{w}_k - w^*\| = \|\bar{w}_k - w_k + w_k - w^*\| \leq \|\bar{w}_k - w^*\| + \|w_k - w^*\|$$

$$\leq \|w_k - w^*\| + \|w_k - w^*\| \leq b_1/2 + b_1/2 = b_1,$$

Means that $\bar{w}_k \in N(w^*, b_1)$. Hence from (2.28) it follows:

$$\|E_k\| \geq c_1 dist(w, W^*)$$

And since from hypothesis (3) we have

$$c_1\|\bar{w}_k - w_k\| \leq \|E_k\|$$

Then from (11) the Levenberg - Marquardt parameter μ_k satisfies:

$$c_1{}^{\delta}\|\bar{w}_k - w_k\|^{\delta} \leq \mu_k = \|E_k\|^{\delta} \leq L_2{}^{\delta}\|\bar{w}_k - w_k\|^{\delta} \quad (14)$$

Define,

$$Q_k(\rho) = \|E_k + J_k\,\rho\|^2 + \mu_k\|\rho\|^2 \quad (15)$$

From (12) it follows that ρ_k is a fixed point of $Q_k(\rho)$. Since,

$\bar{w}_k \in N(w^*, b_1)$ and $b_1 < 1$, we get: $\|\rho_k\|^2 \leq \frac{Q_k(\rho_k)}{\mu_k}$

Since, $w_{k+1} = w_k + \rho_k$

Then, by definition of $Q_k(\rho)$ we have:

$$\|\rho_k\|^2 \leq \frac{Q_k(\bar{w}_k - w_k)}{\mu_k} = \frac{\|E_k + J_k(\bar{w}_k - w_k)\|^2 + \mu_k\|\bar{w}_k - w_k\|^2}{\mu_k}$$

$$= \frac{\|E_k + J_k(\bar{w}_k - w_k)\|^2}{\mu_k} + \|\bar{w}_k - w_k\|^2$$

From (10) and since $\left[c_1{}^{\delta}\|\bar{w}_k - w_k\|^{\delta} \leq \mu_k \quad \Rightarrow \quad \frac{1}{\mu_k} \leq c_1{}^{-\delta}\|\bar{w}_k - w_k\|^{-\delta}\right]$

Then we have

$$\leq L_1^2 {c_1}^{-\delta} \|\overline{w}_k - w_k\|^4 \|\overline{w}_k - w_k\|^{-\delta} + \|\overline{w}_k - w_k\|^2$$

$$\leq L_1^2 {c_1}^{-\delta} \|\overline{w}_k - w_k\|^{4-\delta} + \|\overline{w}_k - w_k\|^2$$

$$\leq \left(L_1^2 {c_1}^{-\delta} + 1\right) \|\overline{w}_k - w_k\|^2$$

The above inequality implies that:

$$\|\rho_k\| \leq \sqrt{(L_1^2 {c_1}^{-\delta} + 1)}\, \|\overline{w}_k - w_k\|$$

From hypothesis (2.3) we get

$$\|\rho_k\| \leq c_2 dist(w_k, W^*)$$

Where $c_2 = \sqrt{L_1^2 {c_1}^{-\delta} + 1}$

Lemma 5.5

If the hypotheses 2 and 3 are satisfied and $w_{k+1}, w_k \in N\left(w^*, \frac{b_4}{2}\right)$ then we have:

$$dist(w_{k+1}, W^*) = dist(w_k + \rho_k, W^*) \leq c_3 dist(w_k, W^*)^{\frac{2+\delta}{2}} \qquad (16)$$

Where $c_3 = \left(\sqrt{L_1^2 + L_2^\delta} + L_1 c_2^2\right) / c_1$

Proof

Since $Q_k(\rho_k) \leq Q_k(\overline{w}_k - w_k)$ and $w_{k+1} = w_k + \rho_k$

$\|E_k + J_k(\overline{w}_k - w_k)\|^2 + \mu_k \|\overline{w}_k - w_k\|^2$ (by definition of $Q_k(\rho_k)$)

$\leq L_1^2 \|\overline{w}_k - w_k\|^4 + \mu_k \|\overline{w}_k - w_k\|^2$ (by (10))

$\leq L_1^2 \|\overline{w}_k - w_k\|^4 + L_2^\delta \|\overline{w}_k - w_k\|^\delta \|\overline{w}_k - w_k\|^2$ (by (14))

$= L_1^2 \|\overline{w}_k - w_k\|^4 + L_2^\delta \|\overline{w}_k - w_k\|^{2+\delta} \leq \left(L_1^2 + L_2^\delta\right) \|\overline{w}_k - w_k\|^{2+\delta}$

By Taylor series we get

$$\|E(w_k + \rho_k)\| = \|E(w_{k+1})\| \leq \|E_k + J_k \rho_k\| + L_1 \|\rho_k\|^2$$

$$\leq \sqrt{Q_k(\rho_k)} + L_1 \|\rho_k\|^2 \quad (\text{by definition of } Q_k(\rho_k))$$

$$\leq \sqrt{L_1^2 + L_2^\delta \, \|\overline{w}_k - w_k\|^{2+\delta}} + L_1 c_2^2 \, \|\overline{w}_k - w_k\|^2 \quad (\text{by (13)})$$

$$\leq \sqrt{L_1^2 + L_2^{\delta}}\, \|\bar{w}_k - w_k\|^{\frac{2+\delta}{2}} + L_1 c_2^2\, \|\bar{w}_k - w_k\|^2$$

$$\leq \left(\sqrt{L_1^2 + L_2^{\delta}} + L_1 c_2^2\right) \|\bar{w}_k - w_k\|^{\frac{2+\delta}{2}}$$

Multiplying both sides of the last inequality by $\frac{1}{c_1}$ we get

$$\frac{1}{c_1}\|E(w_k + \rho_k)\| \leq \frac{1}{c_1}\left(\sqrt{L_1^2 + L_2^{\delta}} + L_1 c_2^2\right) dist(w_k, W^*)^{\frac{2+\delta}{2}}$$

Since, $\|E(w_k + \rho_k)\| \geq c_1 dist(w_k + \rho_k, W^*)$ (by (9))

Then $dist(w_k + \rho_k, W^*) \leq \frac{1}{c_1}\|E(w_k + \rho_k)\|$,

Hence, $dist(w_k + \rho_k, W^*) \leq \frac{1}{c_1}\|E(w_k + \rho_k)\| \leq c_3 dist(w_k, W^*)^{\frac{2+\delta}{2}}$,

Where $c_3 = \frac{\sqrt{L_1^2 + L_2^{\delta}} + L_1 c_2^2}{c_1}$

Theorem 5.6

If the hypotheses (2) and (3) are satisfied and w_0 is chosen to be sufficiently near to W^*, then $w_{k+1} = w_k + \rho_k$ converges super linearly to the solution $\bar{w}$ of equation (7).

Proof

Let $r = \min\left\{\frac{b_1}{2(1+11c_2)}, \frac{1}{2c_3^{2/\delta}}\right\}$. Firstly by induction, we show that:

If $w_0 \in N(w^*, r)$, i.e., $\|w_0 - w^*\| \leq r$ then $w_k \in N\left(w^*, \frac{b_1}{2}\right)$ for all k.

From lemma 2.4, it follows that:

$$\|w_1 - w^*\| = \|w_0 + \rho_0 - w^*\| \leq \|w_0 - w^*\| + \|\rho_0\|$$

$$\leq \|w_0 - w^*\| + c_2\|w_0 - \bar{w}_0\| \quad \text{(by (13))}$$

$$\leq (1 + c_2)r \leq \frac{b_1}{2}$$

Which means $w_1 \in N\left(w^*, \frac{b_1}{2}\right)$. Suppose $w_i \in N\left(w^*, \frac{b_1}{2}\right)$ for $i = 2, \ldots, k$.

Then we have (from Lemma 2.5):

$$\|w_i - \bar{w}_i\| \le c_3\|w_{i-1} - \bar{w}_{i-1}\|^{\frac{2+\delta}{2}} \le \cdots \le c_3^{\left(\frac{2+\delta}{2}\right)^i - 1}\|w_0 - w^*\|^{\left(\frac{2+\delta}{2}\right)^i}$$

$$\le c_3^{\frac{2}{\delta}\left(\left(\frac{2+\delta}{2}\right)^i - 1\right)}\|w_0 - w^*\|^{\left(\frac{2+\delta}{2}\right)^i} \le r\left(\frac{1}{2}\right)^{\left(\frac{2+\delta}{2}\right)^i - 1} = r\left(\left(\frac{1}{2}\right)^{\left(\frac{2+\delta}{2}\right)^i}\left(\frac{1}{2}\right)^{-1}\right)$$

$$= 2r\left(\frac{1}{2}\right)^{\left(\frac{2+\delta}{2}\right)^i} \le 2r\left(\frac{1}{2}\right)^{\left(\frac{3}{2}\right)^i}$$

So, from the definition of r, we get:

$$\|w_{k+1} - w^*\| = \|w_{k+1} - w_k + w_k - w^*\|$$

$$\le \|w_{k+1} - w_k\| + \|w_k - w^*\|$$

$$= \|\rho_k\| + \|w_k - w_{k-1} + w_{k-1} - w^*\|$$

$$\le \|\rho_k\| + \|w_k - w_{k-1}\| + \|w_{k-1} - w^*\|$$

$$= \|\rho_k\| + \|\rho_{k-1}\| + \|w_{k-1} - w_{k-2} + w_{k-2} - w^*\|$$

$$\le \|\rho_k\| + \|\rho_{k-1}\| + \|w_{k-1} - w_{k-2}\| + \|w_{k-2} - w^*\|$$

$$= \|\rho_k\| + \|\rho_{k-1}\| + \|\rho_{k-2}\| + \|w_{k-2} - w^*\|$$

$$\le \|\rho_k\| + \|\rho_{k-1}\| + \|\rho_{k-2}\| + \cdots + \|\rho_{k-(k-1)}\| + \|w_{k-(k-1)} - w^*\|$$

$$\le \|\rho_k\| + \|\rho_{k-1}\| + \|\rho_{k-2}\| + \cdots + \|\rho_1\| + \|w_1 - w^*\|$$

Hence, $\|w_{k+1} - w^*\| \le \|w_1 - w^*\| + \sum_{i=1}^{k}\|\rho_i\|$

$$\le (1 + c_2)r + c_2\sum_{i=1}^{k}\|w_i - \bar{w}_i\|$$

$$\le (1 + c_2)r + 2rc_2\sum_{i=1}^{k}\left(\frac{1}{2}\right)^{\left(\frac{3}{2}\right)^i}$$

$$\le (1 + c_2)r + 2rc_2\left(4 + \sum_{i=1}^{k}\left(\frac{1}{2}\right)^{\left(\frac{3}{2}\right)^i}\right)$$

$$\leq (1+9c_2)r + 2rc_2 \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i$$

$$\leq (1+11c_2)r$$

$$\leq \frac{b_1}{2} \qquad \left(\text{because } r = \frac{b_1}{2(1+11c_2)}\right)$$

So $w_{k+1} \in N\left(w^*, \frac{b_1}{2}\right)$. Now, if w_0 is chosen near to W^*, then all w_k contained in $N\left(w^*, \frac{b_1}{2}\right)$. Then from (16) we get:

$$\sum_{k=0}^{\infty} \text{dist}(w_k, W^*) = \sum_{k=0}^{\infty} \text{dist}(w_{k-1} + \rho_{k-1}, W^*) ,$$

$$\leq c_3 \sum_{k=0}^{\infty} \text{dist}(w_{k-1}, W^*)^{\frac{2+\delta}{2}} < \infty$$

Then

$$\sum_{k=0}^{\infty} \text{dist}(w_k, W^*) < +\infty \qquad \text{from (16))}$$

Which implies, due to Lemma 2.4, that:

$$\sum_{k=0}^{\infty} \|\rho_k\| \leq c_2 \sum_{k=0}^{\infty} \text{dist}(w_k, W^*) < +\infty . \qquad for\ all\ k$$

Thus

$$\sum_{k=0}^{\infty} \|\rho_k\| < +\infty$$

So, the sequence of weights $\{w_k\}$ *is* converges to some point $\bar{w} \in W^*$. It is clear that:

$$\text{dist}(w_k, W^*) \leq \text{dist}(w_k + \rho_k, W^*) + \|\rho_k\|$$

$$\leq c_3\, \text{dist}(w_k, W^*)^{\frac{2+\delta}{2}} + \|\rho_k\| \qquad \text{(from (16))}$$

Then the above inequality implies that:

$$\mathrm{dist}(w_k, W^*) \le \|\rho_k\| + \|\rho_k\| = 2\|\rho_k\| \quad , \forall \text{ large } k. \qquad (17)$$

Thus from (13), (16) and (17) we have:

$$\|\rho_{k+1}\| = O\left(\|\rho_k\|^{\frac{2+\delta}{2}}\right)$$

Hence, $\{w_k\}$ converges to the solution $\bar{w} \in W^*$. Therefore, we have:

$$\lim_{k\to\infty} \frac{\|w_{k+1}-\bar{w}\|}{\|w_k-\bar{w}\|^{\frac{2+\delta}{2}}} = \lim_{k\to\infty} \frac{\|\rho_{k+1}\|}{\|\rho_k\|^{\frac{2+\delta}{2}}} \le \tilde{c}, \tilde{c} \quad \in R^+$$

The above inequality shows $\{w_k\}$ converges super linearly to the solution $\bar{w}$, where $\delta \in [1,2)$, and converges quadratically where $\delta = 2$.

If, $J(w^*)$ is a rectangular matrix or singular, then we use SVD of $J(w^*)$, (i.e. first type of modification), Now, suppose that $\{w_k\}$ converges to $w^* \in W^*$, and take the SVD of the $J(w^*)$, which is:

$$J(w^*) = U^* \Sigma^* V^{*T}$$

$$= U^* \begin{bmatrix} \sigma_1^* & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^* & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} V^{*T} = U_1^* \Sigma_1^* V_1^{*T}$$

Where $\sigma_1^* \ge \sigma_2^* \ge \cdots \ge \sigma_r^* > 0$ and rank $(\Sigma_1^*) = r$.

Suppose that the SVD of $J(w_k)$ and its decomposition has the form:

$$J(w_k) = U_k \Sigma_k V_k^T$$

$$= \left(U_{k,1}, U_{k,2}, U_{k,3}\right) \begin{bmatrix} \sigma_1^{(k)} & & & & & & \\ & \ddots & & & & & \\ & & \sigma_r^{(k)} & & & & \\ & & & \sigma_{r+1}^{(k)} & & & \\ & & & & \ddots & & \\ & & & & & \sigma_{r+q}^{(k)} & \\ & & & & & & 0 \\ & & & & & & & \ddots \\ & & & & & & & & 0 \end{bmatrix} \begin{pmatrix} V_{k,1}^T \\ V_{k,2}^T \\ V_{k,3}^T \end{pmatrix}$$

$$= U_{k,1}\Sigma_{k,1}V_{k,1}^T + U_{k,2}\Sigma_{k,2}V_{k,2}^T + U_{k,3}\Sigma_{k,3}V_{k,3}^T \quad (18)$$

Where $\Sigma_{k,1}, \Sigma_{k,2} > 0$, $\Sigma_{k,3} = 0$, rank $(\Sigma_{k,1}) = r$ and rank $(\Sigma_{k,2}) = q \geq 0$.

Now, we neglecting k in $\Sigma_{k,i}$, $U_{k,i}$, and $V_{k,i}$, i= 1, 2, 3. Consequently, (18) can be written as:

$$J_k = U_1\Sigma_1V_1^T + U_2\Sigma_2V_2^T$$

Now, to prove the quadratic convergence of the modification for LM algorithm (first type).

Lemma 5.7

If the hypothesis 2 is satisfied and $w_k \in \mathrm{N}\left(w^*, \frac{b_1}{2}\right)$, then we have:

(a)$\|U_1U_1^TE_k\| \leq L_2\|w_k - \bar{w}_k\|$;

(b)$\|U_2U_2^TE_k\| \leq 2L_1\|w_k - w^*\|^2$;

(c)$\|U_3U_3^TE_k\| \leq L_1\|w_k - \bar{w}_k\|^2$.

Proof

The result of (a) immediately follows by (11). From the hypothesis 2 (a), and theory of matrix perturbation (see [28]) we get:

$$\|diag(\Sigma_1 - \Sigma_1{}^*, \Sigma_2, 0)\| \leq \|J_k - J^*\| \leq L_1\|w_k - w^*\|$$

From the above relation, we get:

$$\|\Sigma_1 - \Sigma_1{}^*\| \leq L_1\|w_k - w^*\| \text{ and } \|\Sigma_2\| \leq L_1\|w_k - w^*\| \quad (19)$$

Let $s_k = -J_k^+E_k$ (20)

where J_k^+ is representing the pseudo inverse of J_k, and s_k is the least squares solution of min $\|E_k + J_ks_k\|$, now multiply (20) by J_k from left side we get:

$$J_ks_k = -J_kJ_k^+E_k$$

$$\begin{aligned} E_k + J_ks_k &= E_k + J_k(-J_k^+E_k) \\ &= (I - J_kJ_k^+)E_k \\ &= (I - U_1U_1^T - U_2U_2^T)E_k \\ &= U_3U_3^TE_k \end{aligned}$$

Then by taking $\|\cdot\|$ for two sides we have:

$$\|U_3U_3^TE_k\| = \|E_k + J_ks_k\| \leq \|E_k + J_k(\bar{w}_k - w_k)\|$$

$\leq L_1\|w_k - \bar{w}_k\|^2$. The proof (c) is completed.

Now let $\tilde{J}_k = U_1\Sigma_1 V_1^T$, and $\tilde{s}_k = -\tilde{J}_k^+ E_k$. Since s_k is the least squares solution of $\min\|E_k + \tilde{J}_k\tilde{s}_k\|$, from (20) and (19) it follows that:

$$E_k + \tilde{J}_k\tilde{s}_k = E_k + \tilde{J}_k(-\tilde{J}_k^+ E_k)$$
$$= (I - \tilde{J}_k\tilde{J}_k^+)E_k$$
$$= (I - U_1U_1^T)E_k$$
$$= (U_2U_2^T + U_3U_3^T)E_k$$

Then by taking $\|\cdot\|$ for two sides we have:

$$\|(U_2U_2^T + U_3U_3^T)E_k\| = \|E_k + \tilde{J}_k\tilde{s}_k\|$$
$$\leq \|E_k + (J_k - J_k + \tilde{J}_k)(\bar{w}_k - w_k)\|$$
$$\leq \|E_k + J_k(\bar{w}_k - w_k)\| + \|(\tilde{J}_k - J_k)(\bar{w}_k - w_k)\|$$

From (10) we get:

$$\leq L_1\|\bar{w}_k - w_k\|^2 + \|U_1\Sigma_1V_1^T - (U_1\Sigma_1V_1^T + U_2\Sigma_2V_2^T)(\bar{w}_k - w_k)\|$$
$$= L_1\|\bar{w}_k - w_k\|^2 + \|(U_2\Sigma_2V_2^T)(\bar{w}_k - w_k)\|$$
$$\leq L_1\|\bar{w}_k - w_k\|^2 + \|(U_2\Sigma_2V_2^T)\|\|\bar{w}_k - w_k\|$$

From (19) we have:

$$\leq L_1\|\bar{w}_k - w_k\|^2 + L_1\|w_k - w^*\|\|\bar{w}_k - w_k\|$$
$$\leq 2L_1\|w_k - w^*\|^2$$

Hence : $\|(U_2U_2^T + U_3U_3^T)E_k\| \leq 2L_1\|w_k - w^*\|^2$

By the orthogonal property of U_2 and U_3, we get the result of (b).

$\|U_2U_2^T E_k\| \leq 2L_1\|w_k - w^*\|^2$. The proof is completed.

Theorem 5.8

If the hypotheses 2 and 3 are satisfied and the sequence of the weights $\{w_k\}$ which generated by the modified Levenberg-Marquardt algorithm, then $\{w_k\}$ converges quadratically to the solution of equation (7).

Proof

The current iteration for search direction of modified LM algorithm is:

$$\rho_k = -(J_k^TJ_k + \mu_kI)^{-1}J_k^TE_k$$

Apply the SVD of J_k , as: $J_k = U_k \Sigma_k V_k^T$, then we have:

$$\rho_k = -[(U_k \Sigma_k V_k^T)^T (U_k \Sigma_k V_k^T) + \mu_k I]^{-1} (U_k \Sigma_k V_k^T)^T E_k$$

$$= -\left[V_k \Sigma_k^T \underbrace{U_k^T U_k}_{=I} \Sigma_k V_k^T + \mu_k I \right]^{-1} (V_k \Sigma_k^T U_k^T) E_k$$

Since U is orthogonal, i.e., $U_k^{-1} = U_k^T$

$$= -(V_k \Sigma_k^2 V_k^T + \mu_k I)^{-1} (V_k \Sigma_k U_k^T) E_k \quad \{\text{since } \Sigma_k^T = \Sigma_k \, is \, a \text{ diagonal matrix}\}$$

$$= -\left[V_k \Sigma_k^2 V_k^T + \mu_k I \underbrace{(V_k V_k^T)}_{=I} \right]^{-1} (V_k \Sigma_k U_k^T) E_k$$

Since V is orthogonal, i.e., $V_k^{-1} = V_k^T$

$$= -V_k (\Sigma_k^2 V_k^T + \mu_k I V_k^T)^{-1} (V_k \Sigma_k U_k^T) E_k$$

$$= -V_k (\Sigma_k^2 + \mu_k I)^{-1} \underbrace{V_k^T V_k}_{=I} \Sigma_k U_k^T E_k$$

Hence,

$$\rho_k = -V_k (\Sigma_k^2 + \mu_k I)^{-1} \Sigma_k U_k^T E_k \qquad (21)$$

Now when, $J_k = U_1 \Sigma_1 V_1^T + U_2 \Sigma_2 V_2^T$, then we have :

$$\rho_k = -V_1 (\Sigma_1^2 + \mu_k I)^{-1} \Sigma_1 U_1^T E_k - V_2 (\Sigma_2^2 + \mu_k I)^{-1} \Sigma_2 U_2^T E_k \qquad (22)$$

New, we need to prove $\|(w_{k+1} - w^*)\| = o(\|w_k - w^*\|^2)$

According to the Taylor expansion, the objective function (energy function or error function) may be written as follows:

$$E(w_{k+1}) = E(w_k) + E'(w_k)(w_{k+1} - w_k) + o(w_k^2)$$

$$= E_k + J_k \rho_k \qquad \{\text{since } w_{k+1} = w_k + \rho_k\}$$

Now we have:

$$E_k + J_k \rho_k = E_k + (U_1 \Sigma_1 V_1^T + U_2 \Sigma_2 V_2^T) [-V_1 (\Sigma_1^2 + \mu_k I)^{-1} \Sigma_1 U_1^T E_k - V_2 (\Sigma_2^2 + \mu_k I)^{-1} \Sigma_2 U_2^T E_k]$$

$$= E_k - \left[U_1 \Sigma_1 \underbrace{V_1^T V_1}_{I} (\Sigma_1^2 + \mu_k I)^{-1} \Sigma_1 U_1^T E_k \right]$$

$$-\left[\underbrace{U_1\Sigma_1\underbrace{V_1^TV_2}_{o}\ (\Sigma_2^2\ +\mu_kI\ )^{-1}\ \Sigma_2\ U_2^T\ E_k}_{0}\right]$$

$$-\left[\underbrace{U_2\Sigma_2\underbrace{V_2^T\ V_1}_{0}\ (\Sigma_1^2\ +\mu_kI\ )^{-1}\ \Sigma_1\ U_1^T\ E_k}_{0}\right]$$

$$-\left[U_2\Sigma_2\underbrace{V_2^TV_2}_{I}\ \ (\Sigma_2^2\ +\mu_kI\ )^{-1}\ \Sigma_2\ U_2^T\ E_k\right]$$

{ Since V is orthogonal, i.e., $V_k^{-1} = V_k^T$ and $V_k^T\ V_i = 0, \forall\ k \neq i$ }

$= E_k - U_1\Sigma_1\ (\ \Sigma_1^2\ +\mu_kI\ )^{-1}\ \Sigma_1\ U_1^T\ E_k - U_2\Sigma_2\ (\ \Sigma_2^2\ +\mu_kI\ )^{-1}\ \Sigma_2\ U_2^T\ E_k$

By the SVD

$=\mu_k\ U_1\ (\Sigma_1^2\ +\mu_kI\ )^{-1}U_1^T\ E_k + \mu_kU_2\ (\ \Sigma_2^2\ +\mu_kI\ )^{-1}U_2^T\ E_k + U_3U_3^T E_k$ (23)

Since $\{w_k\}$ converges to w^* super linearly, we assume that:

$$L_1\|w_k - w^*\| < \frac{\sigma_r^*}{2}\ ,\qquad \forall \text{ sufficient large } k. \tag{24}$$

Then from (19) we get:

$\|\ (\Sigma_1^2\ +\mu_kI\ )^{-1}\| \leq \|\Sigma_1^{-2}\|$

From (12), we have:

$\|\Sigma_1 - \Sigma_1{}^*\| \leq L_1\|w_k - w^*\|$

$\|\Sigma_1\| + \|\Sigma_1{}^*\| \leq L_1\|w_k - w^*\|$

$\|\Sigma_1\| \leq L_1\|w_k - w^*\| - \|\Sigma_1{}^*\|$

$\|\Sigma_1\|^2 \leq (\|\Sigma_1{}^*\| - L_1\|w_k - w^*\|)^2$

$$\|\Sigma_1\|^{-2} \leq \frac{1}{(\sigma_r^* - L_1\|w_k - w^*\|)^2}$$

From (24), we have:

$$\|\Sigma_1\|^{-2} < \frac{1}{\left(\sigma_r^* - \frac{\sigma_r^*}{2}\right)^2} = \frac{4}{\sigma_r^{*2}} \tag{25}$$

Also, $\Sigma_2^2\ +\mu_kI \geq \mu_k$, and $(\ \Sigma_2^2\ +\mu_kI\ )^{-1} \leq \mu_k^{-1}$

Then:

$$\|(\ \Sigma_2^2\ + \mu_k I\ )^{-1}\| \le \mu_k^{-1} \qquad (26)$$

From the inequalities (25) and (26) together with (14) and Lemma 2.7 we have:

$$\|E_k + J_k\rho_k\| \le \mu_k U_1 \frac{4}{\sigma_r^{*2}}\ U_1^T E_k + \mu_k U_2\, \mu_k^{-1}\, U_2^T E_k + U_3 U_3^T E_k$$

$$\le \frac{4}{\sigma_r^{*2}}\ \mu_k \underbrace{U_1\, U_1^T}_{I}\, E_k + \underbrace{\mu_k \mu_k^{-1}}_{1}\ U_2 U_2^T E_k + U_3 U_3^T E_k$$

(because U is orthogonal , i.e., $U_k^{-1} = U_k^T$)

From (14) we get:

$$\le \frac{4}{\sigma_r^{*2}} L_2{}^{\delta} \|w_k - w^*\|^{\delta}\ L_2\|w_k - w^*\| + (\ U_2 U_2^T\ + U_3 U_3^T) E_k$$

From lemma (2.7) we get:

$$\le \frac{4}{\sigma_r^{*2}} L_2{}^{1+\delta} \|w_k - w^*\|^{1+\delta}\ + 2L_1\|w_k - w^*\|^2$$

$$\le \left(\frac{4}{\sigma_r^{*2}} L_2{}^{1+\delta} + 2L_1\right)\|w_k - w^*\|^2 \qquad ,\delta \in [1,2]$$

Let $c_4 = \frac{4}{\sigma_r^{*2}} L_2{}^{1+\delta} + 2L_1$, then we get

$$\|E_k + J_k\rho_k\| \le c_4\|w_k - w^*\|^2 \qquad (27)$$

From (2.28), we get

$$c_1 dist(w_{k+1}, W^*) \le \|E(w_{k+1})\| = \|E(w_k + \rho_k)\|$$

By Taylor series we have:

$$c_1 dist(w_{k+1}, W^*) \le \|E_k + J_k\rho_k\| + L_1\|\rho_k\|^2$$

From (27) and lemma (2.4) we get

$$\le c_4\|w_k - w^*\|^2 + c_2^2 L_1\|w_k - w^*\|^2$$

$$\le (c_4 + c_2^2 L_1)\|w_k - w^*\|^2$$

It follows from (17) and Lemma 5.4 that

$$\|\rho_{k+1}\| = o(\|\rho_k\|^2)$$

Which implies that the sequence of weight $\{w_k\}$ converges quadratically to w^*, namely,

$$\|(w_{k+1} - w^*)\| = o(\|w_k - w^*\|^2)$$

The proof is completed.

The Flowchart of modified Levenberg - Marquardt algorithm which explains the implementation of the algorithm with the SVD for Jacobian matrix given in Figure (1).

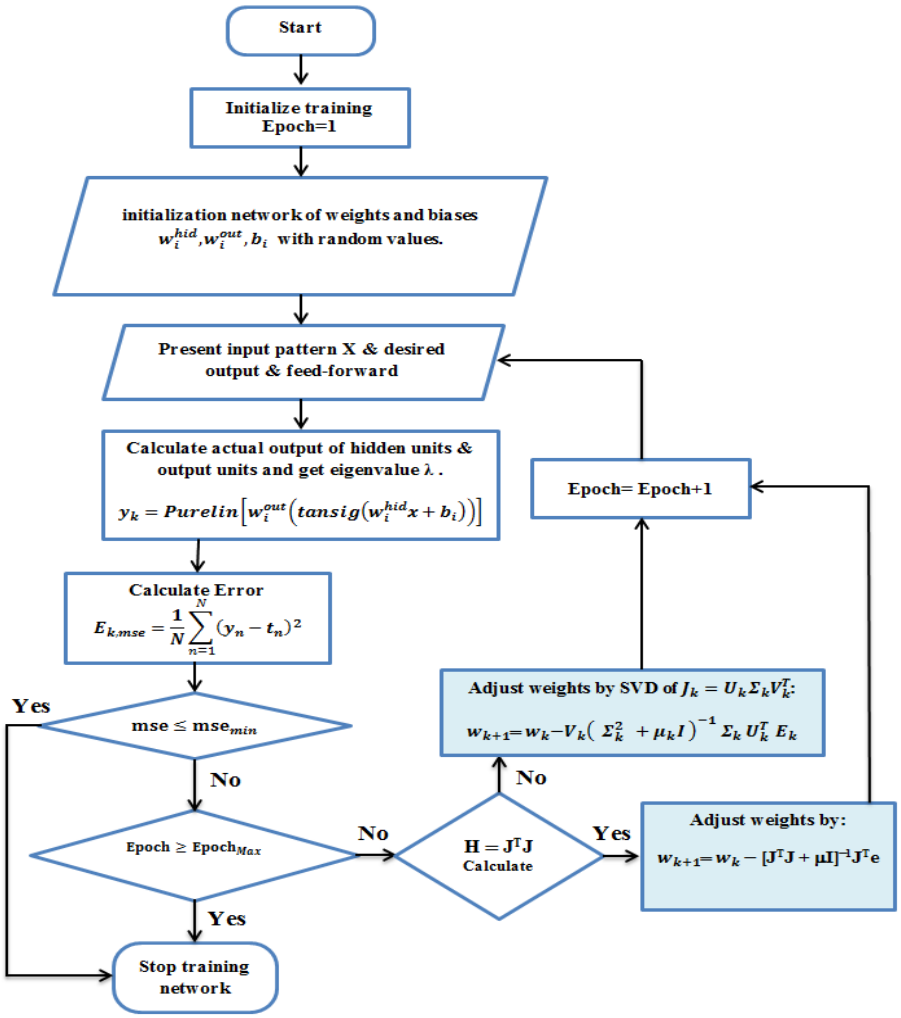

Figure1: Flowchart of modify Levenberg-Marquardt algorithm

References

[1] Haq F. I. , "Numerical Solution Of Boundary-Value And Initial-Boundary-Value Problems Using Spline Functions", Ph.D. Thesis, Faculty of Engineering Sciences GIK Institute of Engineering Sciences and Technology ,Topi , Pakistan, May 2009.

[2] M. Kevin S. , " An Artificial Neural Network Method For Solving boundary Value Problems With Arbitrary Irregular boundaries", Ph.D. thesis in Mechanical Engineering, Georgia Institute of Technology, May, 2006.

[3] Stanevski N. , Tsvetkov D. , "On the Quasi – Newton Training Method for Feed Forward Neural Networks", International Conference on Computer System and Technologies,2004.

[4] Tawfiq L . N. M. , Eqhaar Q. H., "On Multilayer Neural Networks And Its Application For Approximation Problem, 3rd scientific conference of the College of Science ", University of Baghdad. 24 to 26 March ,2009.

[5] Tawfiq L. N. M., "Improving Gradient Descent method For Training Feed Forward Neural Networks", International Journal of Modern Computer Science & Engineering, Vol. 2, No. 1,(2013), pp: 1-25.

[6] T. Valanarasu , N. Ramanujam , "An Asymptotic Initial Value Method For Second Order Singular Perturbation Problems Of Convection-Diffusion Type With A Discontinuous Source Term" ,J. Appl. Math. & Computing, Vol. 23, (2007), No. 1 - 2, pp. 141 – 152.