Stability analysis and Optimization Transcendental Neural Network Learning

Neeraj Sahu
Associate Professor: -
Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore (M.P) India.
Email: - neeraj_maths1@yahoo.co.in

ABSTRACT

The back propagation (B.P) procedure is a useful partial application of weight change in artificial neural networks. Here we use two term algorithms for dynamic learning rate (LR) and the momentum factor (MF). Here disadvantages of these two term BP algorithms are the native minimum and reduce confluence speed and limited real time application. So, we add an additional term named proportional factor (PF) for two terms B.P algorithm. This PF improves the speed of the BP algorithm and decreases the confluence of the B.P algorithm. These criteria are evaluating convergence for required facilities and use of three term BP algorithm. In this paper we define transcendental convergence of three term back propagation algorithm with optimization derivative information. This paper satisfies some conditions for learning parameters of the B.P algorithm. We present learning rate, momentum factor and proportional factor derivative approach. These approaches presented derivative of weight space and using forward and backward procedures.

Keywords -- Back propagation algorithm, Stability analysis, Optimization Technique, Transcendental function, momentum factor (MF), learning rate (LR), proportional factor (PF)

I. INTRODUCTION

1.1 Neural Network

An elementary computing unit of the nervous system is neurons or nervous cells. Humans apparently have 10^{10} and 10^{11} neurons perhaps more. The human nervous system consumes close to 25% of the body's energy and it makes up only 1% or 2% of body weights. It requires far more energy than most tissue [5] this tells us how active this system.

1.2 Historical Background of Neural Network

McCulloch and Pitts established a model for artificial neural network created on simple logic functions such as "X OR Y" and "X AND Y". Rosenblatt [13] stirred substantial attention and action in this field when he invented and created the Perception with three layers, the middle layers and association layer. This structure could connect or associate a assigned input to a random output unit.

In 1960 Windrow and Hoff develop a system called ADALINE (Adaptive Linear Element). Here technique applied for learning was distinct to that of the perception, it is working the Least Mean square (LMS) learning rule. Minsky (1969) wrote a paper and book [8] in which they simplified the limits of single layer Perceptions to multi layered system.

1.3 Type of Neural Network

(a) Artificial Neural Networks

information processing system uses technique of artificial neural network. Here some implementation properties mutual with biological neural networks [7] Artificial neural network has been established simplifications of mathematical models of human thought or neural biology, based on the statements that

(i) Information handling appears at various easy elements called neurons. Every neuron has an inner state, called its beginning or motion level.

(ii) Rays proceed between neurons over connection links.

(iii) Every correlation link has a correlated weight. Which in an average neural network multiplies the signal transferred. Weight represents knowledge being used by the network to explain a problem.

(iv) Every neuron applies an activation function to its net input to determine its output signal.

(v)The technique of neural network applied broad selection of problem, storage, remembering data, optimization, and pattern recognition.

(b) Biological Neural Networks

There are three types of factors of biological neuron dendrites, soma and axon useful and understanding for artificial neurons. Here biological neurons properties propose artificial neural networks [7].

(i) There are many signals received by processing element

(ii) weight and receiving synapses could be useful for modified signal

(iii) Neural transmit a signal output only appropriate circumstance.

(iv) Sums weight input described by processing element

1.4 Uses of Neural Network

(i) Investing analysis

(ii) Sign analysis

(iii) Procedure control

(iv) Monitoring

(v) Advertising

1.5 New application areas

(i) Neural Networks are becoming progressively part of system that are created as a good white toy
(ii) Neural networks are a useful part of soft computing and neural computing.

1.6 Optimization Techniques

(i) Definition- Any problem that requires a positive decision to be made can be classified as operation research (Optimization Technique). The approach used in decision making has changed considerably over the years. The name (O.R) probably came from a program undertaking by Great Britain during World War II "Research in Military Operation."

(ii) Definition - Optimization technique is useful for solving complex real word problems. All engineering and science branches. There are many algorithms designed for technological front by inspiration from different phenomena. Here some admired algorithms named as Genetic Algorithm (GA) based on Darwin's principle of survival of the fittest, Ant Colony Optimization (ACO) based on the foraging behavior of ants, Particle Swarm Optimization (PSO) based on the behavior of birds flocking in swarms and many more. There are many algorithms proposed for technological front by motivation from different phenomena.

(a). linear optimization

(b). Meta-heuristics

(c). Nature inspired Optimization.

1.7 Application area of Optimization Techniques

(i). In mathematical programming method used rigid body dynamics for solving constraint manifold by ordinary differential equation. There are many several nonlinear geometric restrictions i. e “two points must always coincide”, This surface must not infiltrate any other, or "this point must perpetually lie somewhere on this curve”. In this type of problem, linear complementarity problem solves computation contact forces.

(ii). Design problem solve by optimization designs this technique called design optimization. This is a single subset is the engineering optimization and alternative current subset of the field multidisciplinary design optimization. In aerospace engineering several problems solve this technique and this method also applied in cosmology and astrophysics.

(iii). Economic is intently connected to optimization technique that is significant definition linked economics science as the "analysis of human behaviour as a connection between ends and unusual means" with unconventional uses. Recent optimization concept comprises established optimization concept, but they also intersect with game theory and the study of economic equilibria.

(iv). In electrical engineering several applications of optimization method i.e. active filter design, microwave structure, electromagnetics-based design.

(v). Optimization technique generally use in civil engineering. Transportation engineering and construction management and are amid the major division of civil engineering that closely rely on optimization. Here optimization technique solves maximum usual civil engineering problem.

(vi). Operations technique applied stochastic modelling, simulation to assist and expanded decision-making. Progressively, operations research applies programming of stochastic model, decision dynamic that fit to events.

1.8 Stability

The solution of differential equations describes Stability. The trajectory of dynamical system defines initial condition for small perturbations. The heat equation is an example of unchanging partial differential calculation since minor agitation of early data lead to minor variation in temperature after some time results of supreme principle in partial differential equation useful to find distance among L_P norm or sub norm while differential geometry portions the distance among space using the Gromov -Hausdroff distance.

In the dynamical system, if the forward orbit is in a minor neighborhood or it remains in a minor neighborhood called Lyapunov stable. There are different conditions that have been generated to show stability or instability of an orbit. Below satisfactory situations eigenvalues of matrices might be changed to a well-studied difficulty. There are popular methods that include Lyapunov functions. Generally, we can apply any one method stability criteria.

1.9 Studying Method in ANN

(i). Studying method change on the quantity of layers in the network single layers or multi layers.

(ii). Learning process also depends on signal flow of direction i.e feedforward or recurrent neural network.

(iii). There are numeral of node in the input layer is equivalent to the numeral of structures of input data set. The numeral of output node will express in probable outcome i.e. the numeral of modules in the basis of supervised studying and the numeral of hidden layers chosen by the user. Here hidden layer nodes provide higher performance but too many nodes in the hidden layers results in overlapping as well as increasing computation expanse.

(iv). Weight interrelated nodes process the rate of weights involved with every inter correlation among each neuron. So, we can solve many learning problems can be solve correctly, fully a difficult problem such as multi layered feed forward network.

(v). Supervised learning process depends on input variable x and equivalent needed output variable y. Neural networks produce an output created on the input. This output is equated to the needed output. Unsupervised learning process has input figures x and no equivalent output variable. The aim is to structure the primary diagram of the figures to recognize additional figures. Classification and regression are called the keywords of supervised machine learning though are clustering and association.

There are three foremost learning patterns for neural networks: supervised learning, unsupervised learning, and reinforcement learning. Here so many algorithms to train a neural network, including.

(a) Gradient descent

(b) Newton's method

(c) Conjugate gradient

(d) Quasi-Newton method

(e) Levenberg-Marquardt algorithm

II. DESCRIPTION OF PROBLEM

The number of researchers investigates recovering the effectiveness and confluence rate of back propagation system. Though are not higher orders byproducts but determine independent studying rates for every element of weights vector clearly. The conventional B.P require three new parameters for slow convergence rate. In right parameter [1] describe large number of trial run require and proposed a different cost function [14],[15] explain dynamic learning rate momentum factor by derivative information [12] define previous step modification and learning rate though a genetic theorem for self-modification to increase for steepest descent rate. A modern method incremental education for pattern recognition structural adaptation weight adjustment learning systems and apply primary learning to limit the erudition method [6]. The behavior of B.P investigates constant learning rate with static arbitrary input circumstances [3].Two-layer SOM neural network explains theorical basis representing homotopological shape involving input vector and output solution. They independently studied topological organization for DDOA vector and calculated value of AOA in the problem of equivalent linear array [10].

Neural network learning method uses optimization three-term backpropagation system [16]. This method represents optimized Learning Rate, Momentum Factor and Proportional Factor terms and recursive formula evaluating for derivatives and optimization goal with manner Learning Rate, Momentum Factor and Proportional Factor. The behavior of B.P algorithm does not raise computational difficulty. The convergence performance for three backpropagation process [17] established and show the numbers of three term B.P process though verify some condition for stable system and convergence to local minima. Adaptive momentum [4] analyzed with convergence back-propagation algorithm when it uses hidden layer for teaching feedforward neural networks. Convergence theorems describe sufficient conditions for ineffective and effective convergence result. The two objectives accuracy and complexity of the network [2] define hybrid non-subjugated sorting genetic algorithm-II for optimize three-term backpropagation. A multiclass classification problem is effective by experimental results. The stability of RNNs with several equilibria is calculated. There are many recurrent neural network model main factors pretending, many equilibria, activation functions, multi stability and whole stability procedure. The result of total stability and multi stability recurrent neural network [11].

Here we define transcendental function for three-term back propagation algorithm in convergence actions and they satisfy definite situation of coefficient for three-term back propagation algorithm. If the system is stable and covers local minima, then the cost function of local minima is asymptotically stable. cost function also analysis by proportional factor and back propagation algorithm. We also find optimum solution for the learning rate, momentum factor and proportional factor terms. Mentioned equation is modified version Yahya H Zweiri [17].

$$\Delta W(K) = \tanh\alpha\,(-\nabla E(W(K))) + \tanh\beta\,\Delta W(K-1) \quad (1)$$

Let W be a vector established by the whole networks weight and $\nabla E(W(K))$ be the gradient of E at W=W(K) with k=1,2, 3....N existence the iteration number of the weight vector. The momentum term algorithm for two-term back propagation [16] where tanhα learning rate and tanhβ momentum factor correspondingly.

$$\Delta W(K) = -\tanh\alpha\,\nabla E(W(K)) + \tanh\beta\,\Delta W(K-1) + \tanh\gamma\, e(W(K)) \quad (2)$$

We modified this equation for three terms of back propagation algorithm is analyzed. Here we show that the local minima for least square error function are the single nearby asymptotically stable point of algorithm. Then the equation (2) defines as

$$W(K+1) = W(K) - \tanh\alpha\,\nabla E(W(K)) + \tanh\beta\,\Delta W(K-1) + \tanh\gamma\, e(w(K)) \quad (3)$$

$\Psi_1 = W(K)$ and $\Psi_2 = W(K) - W(K-1)$ then equation (3) represents state variable

$$\Psi_1(K+1) = \Psi_1(K) - \tanh\alpha\big(\Psi_1(K)\big) + \tanh\beta\,\big(\Psi_2(K)\big) + \tanh\gamma\, e(\Psi_1(K)) \quad (4)$$

$$\Psi_2(K+1) = -\tanh\alpha\big(\Psi_1(K)\big) + \tanh\beta\,\big(\Psi_2(K)\big) + \tanh\gamma\, e(\Psi_1(K)) \quad (5)$$

Lemma 1. The system of equations (4) and (5) define a equilibrium point at c = (c_1, c_2). If $c_2 = 0$ and $\tanh\alpha\,\nabla E\big(\psi_1(K)\big) = \tanh\gamma\, e\big(W(K)\big)$.

Proof: - Let $\Psi_1(K) = c_1$ and $\Psi_2(K) = c_2$, If $c = (c_1, c_2)$ define equilibrium points

$$\Psi_1(K+1) - \Psi_1(K) = 0 \quad (6)$$

and $\Psi_2(K+1) - \Psi_2(K) = 0$ (7)

when we substitute equation (4) and (5) we find

$$(1 - \tanh\beta)\big(\Psi_2(K)\big) = -\tanh\alpha\,\nabla E\,(\Psi_1(K) + \tanh\gamma\, e(\Psi_1(K)) \quad (8)$$

$$-\tanh\beta\,\big(\Psi_2(K)\big) = -\tanh\alpha\,\nabla E\,(\Psi_1(K)) + \tanh\gamma\, e(\Psi_1(K)) \quad (9)$$

Subtracting eq.(8) from (9) yields $\Psi_2(K) = 0 \Rightarrow c_2 = 0$ replacing $\Psi_2(K) = 0$ in eq.(8) and (9) gives

$$\tanhh\alpha\nabla E\,(\Psi_1(K)) = \tanhh\gamma\, e(\Psi_1(K)) \quad (10)$$

Remark: - If $e(\Psi_1(K)) = 0$ is equilibrium place of equation (4) and (5) so $\nabla E(\Psi_1(K)) = 0$ for c = (c_1 ,c_2) .

The small signal analysis examined regional stability possessions about the equilibrium point (c_1, c_2). Let $\lambda_1 = \Psi_1 - c_1$ and $\lambda_2 = \Psi_2 - c_2$ perturbed signal then we find state equation

$$\lambda_1(K+1) = \lambda_1(K) - \tanh\alpha\nabla E\,(c_1 + \lambda_1(K)) + \tanh\beta(\lambda_2(K) + \tanh\gamma\, e(c_1 + \lambda_1(K)) \quad (11)$$

$$\lambda_2(K+1) = -\tanh\alpha\nabla E\,(c_1 + \lambda_1(K)) + \tanh\beta(\lambda_2(K) + \tanh\gamma\, e(c_1 + \lambda_1(K)) \quad (12)$$

When we can linearize about the equilibrium point c equation (11) and (12) suited

$$\lambda_1(K+1) = \lambda_1(K) - \tanh\alpha\nabla^2 E\,(c_1)\,(\lambda_1(K)) + \tanh\beta(\lambda_2(K) + \tanh\gamma\, e(c_1) - \lambda_1(K)) \quad (13)$$

$$\lambda_2(K+1) = -\tanh\alpha\nabla^2 E\,(c_1)\,(\lambda_1(K)) + \tanh\beta(\lambda_2(K) + \tanh\gamma\, e(c_1) - \lambda_1(K)) \quad (14)$$

If Q is a size of weight vector, then Hessian matrix equivalent to $A \in R^{QXQ}$ and $\nabla e\,(c_1)$ equivalent to $D \in R^{Q\,X\,Q}$

$$\begin{bmatrix}\lambda_1(K+1)\\ \lambda_2(K+1)\end{bmatrix} = \begin{bmatrix}1 - \tanh\alpha\, A + \tanh\gamma\, D & \tanh\beta\, I\\ -\tanh\alpha\, A + \tanh\gamma\, D & \tanh\beta\, I\end{bmatrix}\begin{bmatrix}\lambda_1(K)\\ \lambda_2(K)\end{bmatrix} \quad (15)$$

The more compact from of above matrix

$$\lambda(K+1) = \xi\,\lambda(K) \quad (16)$$

we know that equation (16) is discrete time system and asymptotically stable if ξ has distinct eigen values values $\emptyset_i$ of ξ satisfy by (Leigh, 1985)

$$|\emptyset_1| < 1 \quad (17)$$

Lemma 2: - $\left(\frac{A}{\tanh\gamma} - \frac{D}{\tanh\alpha}\right)$ the corresponding eigen value of λ_i of F, of pairs ξ are given by the solution of quadric equation

$$\emptyset_i^2 - (1 + \tanh\beta - \tanh\alpha\tanh\gamma)\emptyset_i + \tanh\beta = 0 \quad (18)$$

Proof. Whichever A and D invertible by ξ as long as $\tanh\beta \neq 0$. Let ξ eigenvalue is $\emptyset_i$ and nonsingular ξ is nonzero. Let z = (x,y) be non-zero eigen vector analogous to $\emptyset_i$ then

$$\xi\, z = \emptyset_i\, z \quad (19)$$

which directs to

$$x - \tanh\alpha\, Ax + \tanh\gamma\, Dx + \tanh\beta\, y = \emptyset_i x \quad (20)$$

and

$$-\tanh\alpha\, Ax + \tanh\gamma\, Dx + \tanh\beta\, y = \emptyset_i y \quad (21)$$

By substituting equation (21) in equation (20) and resolving for $y(\emptyset_i \neq 0)$ given

$$y = \frac{(\emptyset_i - 1)}{\emptyset_i} x \quad (22)$$

by substituting equation (22) in equation (21) gives

$$(-\tanh\alpha\, A + \tanh\gamma\, D)\, x + ((\emptyset_i - 1) - \frac{\tanh\beta(\emptyset_i - 1)}{\emptyset_i})x \quad (23)$$

Since $\left[\left(\frac{A}{\tanh\gamma}\right) - \left(\frac{D}{\tanh\alpha}\right)\right] = F$ substituting in equation (23) gives

$$F_X = \left(\frac{(\emptyset_i - 1)(\frac{\tanh\beta}{\emptyset_i} - 1)}{\tanh\alpha\tanh\gamma}\right) x \quad (24)$$

Horn & Johnsm 1985 says if vector x fulfilled these equations, then x is called eigen vector of F.

Where $\left(\frac{(\emptyset_i - 1)(\frac{\tanh\beta}{\emptyset_i} - 1)}{\tanh\alpha\tanh\gamma}\right)$ is scalar and nonzero.

Now λ_i eigen value of F and Fx = λ_i x then linear

$$\lambda_i = \left(\frac{(\emptyset_i - 1)(\frac{\tanh\beta}{\emptyset_i} - 1)}{\tanh\alpha \tanh\gamma} \right) \quad (25)$$

With corresponding eigen vector x. Rearranging equation (25) yields

$$\emptyset_i^2 - (1 + \tanh\beta - \tanh\alpha\, \lambda_i \tanh\gamma)\emptyset_i + \tanh\beta = 0 \quad (26)$$

Theorem 1: - If system describe stable condition of equation (13) and (14) and $\emptyset_i$ is roots of equation (26) and satisfy $|\emptyset_i| < 1$. If they satisfy following condition

$$0 < \tanh\beta < 1 \quad \text{and} \quad (27)$$

$$0 < \tanh\alpha \tanh\gamma\, \lambda_i < 4 \quad (28)$$

Proof: - The equation (26) represents the polynomial of second degree. Then equation shows as $f(z) = a_2 + z^2 + a_1 z + a_0 = 0$ from the jury test the roots of f(z) describe a unit circle for the following condition

$$|a_0| < a_2 \quad (29)$$

$$f(1) > 0 \quad (30)$$

$$\text{and } (-1)^2 f(-1) > 0 \quad (31)$$

applying the jury test to equation (26) yields roots within the unit circle if

$$|\tanh\beta| < 1 \quad (32)$$

$$(1 + \tanh\beta) > ((1 + \tanh\beta) - \tanh\alpha \tanh\gamma\, \lambda_i) \quad (33)$$

and

$$(1 + \tanh\beta) > (-(1 + \tanh\beta) + \tanh\alpha \tanh\gamma\, \lambda_i) \quad (34)$$

Hence inequality (33) leads to

$$\tanh\alpha \tanh\gamma\, \lambda_i > 0 \quad (35)$$

and inequality (34)

$$\tanh\beta > \frac{\tanh\alpha \tanh\gamma\, \lambda_i}{2} - 1 \quad (36)$$

since the momentum factor $\tanh\beta$ is positive $0 < tanh\beta < 1$ (37)

using inequality (36) on equation (37) yields

$$0 < \tanh\alpha \tanh\gamma\, \lambda_i < 4 \quad (38)$$

$\tanh\alpha$ and $\tanh\gamma$ values must be positive though accept the system and understand.

III. Inference of optimal Term learning rate, momentum factor and proportional factor

Suppose a group of teaching example pairs describe as (I_1, T_1) (I_2, T_2) ---------- (I_n, T_n) where I_s, $1 \leq s \leq n$ indicate the s^{th} input and T_s, $1 \leq s \leq n$ is the corresponding desired output of back propagation system for multi-layer neural system for random hidden layers then least square error function

$$E = \frac{1}{n\, Z_M} \sum_{S=1}^{n} [\ T_S - O_S^M]^T\ [\ T_S - O_S^M] \tag{39}$$

Where O_S^M, I_S and Z_M is the output vector, input, and output neurons for M-layered network. If feed forward calculation of system with I_S shows input layer

$$O_{S,i}^m = f([W_i^m(K+1)]^T\ O_S^{m-1}) \tag{40}$$

Where $O_{S,i}^m$, $1 \leq i \leq Z_m$ describe the i^{th} output of layer m, $1 \leq m \leq M$ and f(.) is the activation function. If $W_i^m(K+1)$ is a sub vector of W (K+1) then it contains all weights of neurons of layer m-1 to $O_{S,i}^m$ and $O_{S,i}^{m-1}$ is a vector created by all the output of layer m-1and is given by

$$O_{S,i}^m = \begin{cases} [\ 1\ \ O_{S,1}^{m-1} \ldots\ldots O_{S,Z_m}^{m-1}\ ]^T & \text{for } m > 1, \\ [\ 1 \quad I_S^T\ \ ]^T & \text{for } m = 1 \end{cases} \tag{41}$$

Here w(K+1) minimize E is needed for optimization $\tanh\alpha, \tanh\beta, \tanh\gamma$.Here we use as a function of equation with a three independent variables E($\tanh\alpha$, $\tanh\beta$, $\tanh\gamma$) called objective function E.

$$W(K+1) = W(K) + \tanh\alpha\ P(K) + \ \tanh\beta\ \Delta w(K-1) + \ \tanh\gamma\ e\ (w(K)) \tag{42}$$

where $P(K) = -\nabla E\ W(K)$ is descent directional vector exchanging eq. (42) into eq. (40) provides.

$$O_{S,i}^m = f([\ W_i^m(K) + \tanh\alpha P_i^m(K) + \tanh\beta\ \Delta W_i^m(K-1) + \tanh\gamma\ e_i^m(K)]^T\ O_S^{m-1}) \tag{43}$$

By the calculation of first and second derivative of E with respect to $\tanh\alpha, \tanh\beta, \tanh\gamma$ yield

$$g(\tanh\alpha, \tanh\beta, \tanh\gamma) = \begin{bmatrix} \frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial \tanh\alpha} \\ \frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial \tanh\beta} \\ \frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial \tanh\gamma} \end{bmatrix} \tag{44}$$

$$\frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial \tanh\alpha} = -\frac{2}{n z_m} \sum_{s=1}^{n} \left[T_S - O_S^M\right] \frac{\partial O_S^M}{\partial \tanh\alpha} \tag{45}$$

$$\frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial \tanh\beta} = -\frac{2}{n z_m} \sum_{s=1}^{n} \left[T_S - O_S^M\right] \frac{\partial O_S^M}{\partial \tanh\beta} \tag{46}$$

$$\frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial \tanh\gamma} = -\frac{2}{n z_m} \sum_{s=1}^{n} \left[T_S - O_S^M\right] \frac{\partial O_S^M}{\partial \tanh\gamma} \tag{47}$$

Hessian matrix of E is defined as

$$H(\tanh\alpha, \tanh\beta, \tanh\gamma) = \begin{bmatrix} \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\alpha)^2} & \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\alpha)\,\partial(\tanh\beta)} & \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\alpha)\,\partial(\tanh\gamma)} \\ \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\beta)\,\partial(\tanh\alpha)} & \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\beta)^2} & \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\beta)\,\partial(\tanh\gamma)} \\ \frac{\partial E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\gamma)\,\partial(\tanh\alpha)} & \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\gamma)\,\partial(\tanh\beta)} & \frac{\partial^2 E(\tanh\alpha,\tanh\beta,\tanh\gamma)}{\partial(\tanh\gamma)^2} \end{bmatrix} \tag{48}$$

When we compute equation (44) for gradient vector and derivatives of O_S^M at $(\tanh\alpha_0, \tanh\beta_0, \tanh\gamma_0)$ for Hessian matrix equation and computed equation (48) thus the derivative of objective function E(X) define as

$$X = [\tanh\alpha, \tanh\beta, \tanh\gamma]^T \tag{49}$$

We use second order and second-degree Taylor polynomial for estimated E(X) and X near $(\tanh\alpha_0, \tanh\beta_0, \tanh\gamma_0)$. The condition of E(X) has continuous second order partial derivative though define as

$$\begin{aligned} E(X) = {} & E(\tanh\alpha_0, \tanh\beta_0, \tanh\gamma_0) + (\tanh\alpha - \tanh\alpha_0)\frac{\partial E}{\partial \tanh\alpha} + (\tanh\beta - \tanh\beta_0)\frac{\partial E}{\partial \tanh\beta} \\ & + (\tanh\gamma - \tanh\gamma_0)\frac{\partial E}{\partial \tanh\gamma} + \frac{1}{2}(\tanh\alpha - \tanh\alpha_0)^2\frac{\partial^2 E}{\partial \tanh^2\alpha} + \frac{1}{2}(\tanh\beta - \tanh\beta_0)^2\frac{\partial^2 E}{\partial \tanh^2\beta} \\ & + \frac{1}{2}(\tanh\gamma - \tanh\gamma_0)^2\frac{\partial^2 E}{\partial \tanh^2\gamma} + (\tanh\alpha - \tanh\alpha_0)(\tanh\beta - \tanh\beta_0)\frac{\partial^2 E}{\partial \tanh\alpha\, \partial \tanh\beta} \\ & + (\tanh\beta - \tanh\beta_0)(\tanh\gamma - \tanh\gamma_0)\frac{\partial^2 E}{\partial \tanh\beta\, \partial \tanh\gamma} + (\tanh\gamma - \tanh\gamma_0)(\tanh\alpha - \\ & \tanh\alpha_0)\frac{\partial^2 E}{\partial \tanh\gamma\, \partial \tanh\alpha} \\ = {} & \frac{1}{2}\Psi^T H_e \Psi + \Psi^T g_e + a_e \end{aligned} \tag{50}$$

Where $\Psi = [\tanh\alpha - \tanh\alpha_0\ \tanh\beta - \tanh\beta_0\ \tanh\gamma - \tanh\gamma_0\ ]^T$, $\eta_0 = E(\tanh\alpha_0, \tanh\beta_0, \tanh\gamma_0)$.

Here equation (44) defines gradient vector g and equation (48) defines hessian matrix H.

Case I: - According to M.A. Wolfe [9] E(X) represents convex set C continuous second partial derivatives and we assume the Hessian matrix H(X) at x for all X in C be positive definite. Here crucial point y of E(X) in C. So, we can say if E(X) is closely convex in C then y is powerful global minimize of E(X) above C. We assume function E (0,0,0) = 0 and gradient E(0,0,0) = 0 for equation (50) and simplifies quadratic polynomial

The discriminates are

$$D_1 = 4\left(\frac{1}{4}E_{\tanh\alpha\tanh\alpha}\right)\left(\frac{1}{4}E_{\tanh\beta\tanh\beta}\right) - \left(E_{\tanh\alpha\tanh\beta}\right)^2$$

$$D_2 = 4\left(\frac{1}{4}E_{\tanh\alpha\tanh\alpha}\right)\left(\frac{1}{4}E_{\tanh\gamma\tanh\gamma}\right) - \left(E_{\tanh\alpha\tanh\gamma}\right)^2$$

$$D_3 = 4\left(\frac{1}{4}E_{\tanh\beta\tanh\beta}\right)\left(\frac{1}{4}E_{\tanh\gamma\tanh\gamma}\right) - \left(E_{\tanh\beta\tanh\gamma}\right)^2$$

$$\begin{aligned} E(X) = {} & \frac{1}{2}(\tanh\alpha)^2 E_{\tanh\alpha\tanh\alpha} + \frac{1}{2}(\tanh\beta)^2 E_{\tanh\beta\tanh\beta} + \frac{1}{2}(\tanh\gamma)^2 E_{\tanh\gamma\tanh\gamma} \\ & + \tanh\alpha\tanh\alpha E_{\tanh\alpha\tanh\alpha} + \tanh\beta\ \tanh\beta\ E_{\tanh\beta\tanh\beta} + \tanh\gamma\tanh\gamma E_{\tanh\gamma\tanh\gamma} \end{aligned} \tag{51}$$

If $E_{\tanh\alpha\tanh\alpha} > 0$, $D_1 > 0$ is a positive definite for H then symmetric matrix and $(D_2 > 0, D_3 > 0)$ the optimal learning rate, momentum factor and proportional factor terms can be calculated as

$$\frac{dE}{d\Psi} = H\Psi + g = 0 \Rightarrow \Psi = -H^{-1}g \tag{52}$$

It is noted that this process defines equation (50) is minimized.

Case II: - If one of D_2 or D_3 is negative and H is a positive definite matrix, then $E(\tanh\alpha, \tanh\beta, \tanh\gamma)$ can't be categorized as convex. Though, $E(\tanh\alpha, \tanh\beta, 0)$ is convex and optimal learning rate and momentum factor terms can be designed as in case first by location $\tanh\gamma = 0$

Case III: - If $E_{\tanh\alpha\tanh\alpha} > 0$ and H is a non-positive definite matrix then the expansion of second order E(tanhα,0,0) convex alongside the descent direction of P(K). We calculate the optimal learning rate in case first by location $\tanh\beta = \tanh\gamma = 0$

Case IV: - Suppose H is non positive definite matrix and $E_{\tanh\alpha\tanh\alpha} < 0$ the optimization aim comportment accelerated declined method along the descent direction P(k) because both $E_{\tanh\alpha}$ and $E_{\tanh\alpha\tanh\alpha}$ accept negative values. Yu and several authors [15] represent the optimal LR and estimated line search method and efficient of supplying an effective descent to the optimization aim.

IV Estimate of sigmoid nonlinear transcendental function

Suppose the equation

$$y = f([\, W_i^M(k) + \tanh\alpha P_i^M(k) + \tanh\beta\, \Delta W_i^M(k-1) + \tanh\gamma\, e_i^M(k)]^T\, O_S^{M-1})$$

Here we describe sigmoidal nonlinear function, for output layer and approximated. Set of liner function

$$f(y) = \begin{cases} m_1 y + b_1 & \text{for } y_1 \le y \le y_2 \\ m_2 y + b_2 & \text{for } y_1 \ge y \\ m_2 y + (2b_1 - b_2) & \text{for } y_1 \le y \end{cases} \tag{53}$$

$$O_S^M = f([\, W_i^k(k) + \tanh\alpha P_i^M(k) + \tanh\beta\, \Delta W_i^m(k-1) + \tanh\gamma\, e_i^m(k)]^T\, O_S^{M-1}) \tag{54}$$

On substituting eq. (54) into equation (45)-(47) and equating $e_{\tanh\alpha}$, $e_{\tanh\beta}$, $e_{\tanh\gamma}$ to zero yield

$$\tanh\alpha\, m_j\, P_i^M \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^T O_s^{M-1} + \tanh\beta\, m_j\, \Delta W_i^M(K-1) \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^T O_s^{M-1}$$
$$+ \tanh\gamma\, m_j\, e_i^M \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^T O_s^{M-1} = \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^T (T_s - m_j W_i^M(k) O_s^{M-1} - b_j) \tag{55}$$

$$\tanh\alpha\, m_j\, P_i^M \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^T O_s^{M-1} + \tanh\beta\, m_j\, \Delta W_i^M(K-1) \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^T O_s^{M-1}$$
$$+ \tanh\gamma\, m_j\, e_i^M \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^T O_s^{M-1} = \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^T (T_s - m_j W_i^M(k) O_s^{M-1} - b_j) \tag{56}$$

$$\tanh\alpha\, m_j\, P_i^M \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^T O_s^{M-1} + \tanh\beta\, m_j\, \Delta W_i^M(K-1) \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^T O_s^{M-1}$$
$$+ \tanh\gamma\, m_j\, e_i^M \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^T O_s^{M-1} = \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^T (T_s - m_j W_i^M(k) O_s^{M-1} - b_j) \tag{57}$$

From equation (59) define a non-singular matrix A_2 then the optimal $\tanh\alpha, \tanh\beta$ and $\tanh\gamma$ can be calculated by solving equation (55)-(57) simultaneously

$$\tau = A_2^{-1} R_2 \tag{58}$$

$$A_2 = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \tag{59}$$

Where

$$A_{11} = \mathrm{m_j}\ \mathrm{P_i^M}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^{T}\ \mathrm{O_s^{M-1}}\ , \qquad A_{12} = \mathrm{m_j}\ \Delta \mathrm{W_i^M(K-1)}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^{T}\ \mathrm{O_s^{M-1}}$$

$$A_{13} = \mathrm{m_j}\ \mathrm{e_i^M}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^{T}\ \mathrm{O_s^{M-1}}\ , \qquad A_{21} = \mathrm{m_j}\ \mathrm{P_i^M}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^{T}\ \mathrm{O_s^{M-1}}$$

$$A_{22} = \mathrm{m_j}\ \Delta \mathrm{W_i^M(K-1)}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^{T}\ \mathrm{O_s^{M-1}}\ , \qquad A_{23} = \mathrm{m_j}\ \mathrm{e_i^M}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^{T}\ \mathrm{O_s^{M-1}}$$

$$A_{31} = \mathrm{m_j}\ \mathrm{P_i^M}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^{T}\ \mathrm{O_s^{M-1}}\ , \qquad A_{32} = \mathrm{m_j}\ \Delta \mathrm{W_i^M(K-1)}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^{T}\ \mathrm{O_s^{M-1}}$$

$$A_{33} = \mathrm{m_j}\ \mathrm{e_i^M}\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^{T}\ \mathrm{O_s^{M-1}}$$

and

$$R_2 = \begin{bmatrix} \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\alpha}\right]^{T}\ (\mathrm{T_s} - \mathrm{m_j W_i^M(k) O_s^{M-1}} - \mathrm{b_j}) \\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\beta}\right]^{T}\ (\mathrm{T_s} - \mathrm{m_j W_i^M(k) O_s^{M-1}} - \mathrm{b_j}) \\ \sum_{s=1}^{n}\left[\frac{\partial O_s^M}{\partial \tanh\gamma}\right]^{T}\ (\mathrm{T_s} - \mathrm{m_j W_i^M(k) O_s^{M-1}} - \mathrm{b_j}) \end{bmatrix} \tag{60}$$

V. CONCLUSIONS

Here we describe ascertains necessary and sufficient condition for confluence and stability actions of three-term back propagation transcendental function equation (35) and (36) satisfy the concurrent of three-term back propagation system. This equation also shows a stable system and will cover local minima. Constraint (36) also defined the large eigen value of matrix F. The most of all cases minima though are sit inside a bounded set because F is bounded and hence if $\tanh\alpha$, $\tanh\gamma$ are adequately small, all the neighboring minima stable. If the system is unstable, it means one eigen value of matrix F is minus. It is also describing cost function of all minima are single locally asymptotically point for the system. This paper shows an optimization approach for development, finds optimal training limits, improving learning rate for three term of back propagation algorithm. We use an optimization approach for transcendental function and generate sigmoidal nonlinearity function.

VI. REFERENCES

1. A.O., Ooyen & B. Neinnhuis "Improving the convergence of the backpropagation algorithm" Neural Networks 1(4): 295-307, 1998.
2. Ashraf Osman Ibrahim, Siti Mariyam Shamsuddin, Nor Bahiah Ahmad, Mohd Najib Mohd Salleh "Hybrid NSGA-II of Three-Term Backpropagation network for multiclass classification problems" IEEE *Xplore*: 31 July 2014
3. C. M. Kuan & K. Hornik "Convergence of learning algorithm with constant learning rates" IEEE Transaction on Neural Networks, 2(5): 484-489.1991.
4. Hongmei Shao, Gaofeng Zheng "Convergence analysis of a back-propagation algorithm with adaptive momentum" Neurocomputing Volume 74 Issue 5, Pages 749-752, February 2011.
5. James A. Anderson. An Introduction to Neural Network: Prentice-Hall of India (P) Ltd.
6. L.M. Fu H.H. Hsu & C.J "Incremental backpropagation learning network". IEEE Transaction on Neural Network, 7(30; 757-761,1996
7. Laurence Fausett. "Fundamentals of Neural Networks Architecture, Algorithms and Application" Pearson Education, Low Price Edition, 1993.
8. Minsky and Papert Perceptron: An introduction to computation geometry, MIT press, expend edition,1969.
9. M.A. Wolfe, "Numerical methods for unconstrained optimization". VNR, Wokingham, U.K, 1978.
10. Pin Wang, Peng Wang, and En Fan "Neural Network Optimization Method and Its Application in Information Processing" Hindawi Mathematical Problem in Engineering Article ID 6665703, 10 pages, doi.org/10.1155/2021/6665703 Volume 2021
11. Peng Liu, Jun Wang and Zhigang Zeng "An Overview of the Stability Analysis of Recurrent Neural Networks with Multiple Equilibria" IEEE Transactions on Neural Networks and Learning system, Vol 34, No. 3 March 2023.
12. R. Salomon & J. L. Hemmen "Accelerating backpropagation through dynamic self-adaption". Neural Networks, 9(4): 589-601,1996.
13. Rosenblatt, Frank. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. In, Psychological Review, Vol. 65, No. 6, pp. 386-408, Nov.1958.
14. X.H., Yu & G.A. Chen "Efficient backpropagation learning using optimal learning rate and momentum" Neural networks, 10(3): 517-527, 1997.
15. X.H., Yu & G.A. Chen, & S.X. Cheng "Dynamic learning rate optimization of the backpropagation algorithm". IEEE Transaction on Neural Network, 6(3):669-677,1995.
16. Yahya H Zweiri "Optimization of a three-term backpropagation algorithm used for neural network learning" International journal of computational intelligence 3(4) 322-327, 2007.
17. Yahya H Zweiri, Lakmal D. Seneviratne, Kaspar Althoefer "Stability analysis of three backpropagation algorithm" Neural Network 18(10) 1341-1347, 2005.