CONVERGENCE ANALYSIS OF NEURAL NETWORKS TRAINING BASED ON STEEPEST DESCENT METHOD

L. Maxnist[1], A. Doudkin[2], V. Golovko[1]
[1]Brest State Technical University, 256 Moskovskaya, Brest, gva@bstu.by
[2]United Institute of Informatics Problems, National Academy of Sciences of Belarus,
6 Surganov str., Minsk, 220012, Belarus, doudkin@newman.bas-net.by

A convergence of neural networks training is investigates in this paper. For training we consider a steepest descent method under the following suppositions: mean square error function is Lipschitz continuous. Two new sufficient conditions are derived to ascertain the neural network training convergence.

Introduction

Let's consider a neural network consisting from *n* neural elements of a distributive layer and *m* neural elements of a target layer. It maps input vectors $\overline{x^k} = \left(x_1^k, \ldots, x_n^k\right)$ to output vectors $\overline{y^k} = \left(y_1^k, \ldots, y_n^k\right) \left(k = \overline{1, L}\right)$ according to the formula $y_j^k = F\left(S_j^k\right)$, where $S_j^k = \sum_{i=1}^{n} w_{ij} x_i^k - T_j, \;\; j = \overline{1, m}, \;\; k = \overline{1, L}$, F is an activation (squashing) function, $w_{ij} \left(i = \overline{1, n}, \;\; j = \overline{1, m}\right)$ are weights connecting neurons of the distributive layer to the neurons of the target layer, $T_j \left(j = \overline{1, m}\right)$ is a bias of neural element *j*.

A key feature of neural networks is that dependence between their inputs and outputs is calculated in a training process. There are two approaches to training - supervised and unsupervised. Different types of networks use different types of training, but supervised training is applied more often and now we shall consider this method. It involves a mechanism of providing the network with the desired output either by manually "grading" the network's performance or by providing the desired outputs with the inputs. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust network parameters (weights and biases) which control the network. This process occurs over and over as the parameters are ever refined. As a rule a performance function for training is mean square error - the average squared error between the network outputs y_j^k and the target outputs t_j^k.

There are two different ways, in which the gradient descent algorithm can be implemented: incremental mode and batch mode. In the incremental mode, the gradient is computed and the weights are updated after each input is applied to the network. In batch mode the network parameters are updated only after the entire training set has been applied to the network (all of the inputs are applied to the network before the parameters are updated). The gradients calculated at each training example are added together to determine the change in the weights and biases. These two methods are often too slow for practical problems. Two main approaches are known to increase higher performance of training algorithms. The first one uses heuristic techniques, which were developed from an analysis of the performance of the standard steepest descent algorithm (the momentum technique, variable learning rate backpropagation, resilient backpropagation etc.) [1]. The second approach based on standard numerical optimization techniques for neural network training: conjugate gradient, quasi-

Newton and Levenberg-Marquardt etc. [2]. At that, considerable efforts are devoted to the analysis of the training convergence [3-5]. Absolutely stable neural networks are desirable in optimization, signal processing and pattern recognition, which have been widely studied in theory and application [1,6].

In this paper we discuss steepest descent method under different assumption about mean square error function. Two new sufficient conditions are derived to ascertain the neural network training convergence, based on the supposition of the Lipschitz continuous of the error functions.

1. Problem definition

Lets $\overline{x^k} = \left(x_1^k, \ldots, x_n^k\right) \left(k = \overline{1,L}\right)$ are input patterns from a training set. The training task of a neural network with a fixed activation function F consists in finding of weight factors $w_{ij} \left(i = \overline{1,n},\ j = \overline{1,m}\right)$ and biases of neural elements $T_j \left(j = \overline{1,m}\right)$, which minimize summary network error E_S given as a deviation of target values t_j^k from corresponding values y_j^k of each *j-th* neuron of the network for *k-th* pattern. As an error of the network it is possible to consider mean- square deviation $E_S = \frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{m}\left(y_j^k - t_j^k\right)^2$, which we shall name a network error function.

A column $\overline{W} = \left(w_{11}, w_{21}, \ldots, w_{n1}, T_1, w_{12}, w_{22}, \ldots, w_{n2}, T_2, \ldots, w_{1m}, w_{2m}, \ldots, w_{nm}, T_m\right)^T$ we shall name an approximate solution or simple a solution of the set of equations by mean deviation method $F\left(\sum_{i=1}^{n} w_{ij} x_i^k - T_j\right) = t_j^k,\ \ j = \overline{1,m},\ \ k = \overline{1,L}$, if $E_S = \frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{m}\left(F\left(\sum_{i=1}^{n} w_{ij} x_i^k - T_j\right) - t_j^k\right)^2$ reaches the least value. It is possible to apply various gradient methods to find the solution, for example, the steepest descent method, a method of the conjugate gradients and their modifications [1, 2].

Let's consider a method of steepest descent

$$\overline{W}(t+1) = \overline{W}(t) - \alpha(t) \cdot \nabla E_S\left(\overline{W}(t)\right) \tag{1}$$

for minimization of the network error function $E_S\left(\overline{W}\right) = \frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{m}\left(F\left(\sum_{i=1}^{n} w_{ij} x_i^k - T_j\right) - t_j^k\right)^2$.

The behaviour of this method will be disused under various suppositions about $E_S\left(\overline{W}\right)$ and $\alpha(t)$.

2. Lipschitz conditions for convergence

Two theorems are proofed below, those are defined the convergence conditions.

Theorem 1. If a gradient of the function $E_S\left(\overline{W}\right)$ satisfies to Lipschitz condition

$$\left\|\nabla E_S\left(\overline{W}\right)-\nabla E_S\left(\overline{V}\right)\right\| \le L\left\|\overline{W}-\overline{V}\right\|,\ \forall \overline{W},\overline{V}\in \Re^{nm+m},\quad L>0, \tag{2}$$

where $\Re^{nm+m}$ denotes *(nm+m)*- dimensional Euclidean space, and $\alpha(t)$ satisfies to the condition

$$0<\varepsilon<\alpha(t)<\frac{2}{L}-\varepsilon,\ \forall t, \tag{3}$$

then the gradient in (1) aspires to zero: $\lim\limits_{t\to\infty}\nabla E_S\left(\overline{W}(t)\right)=0$, and function $E_S\left(\overline{W}\right)$ decreases monotonically : $E_S\left(\overline{W}(t+1)\right)\le E_S\left(\overline{W}(t)\right)$.

Proof. Considering the relation

$$E_S\left(\overline{W}+\overline{V}\right)=E_S\left(\overline{W}\right)+\int_0^1\left(\nabla E_S\left(\overline{W}+\tau\overline{V}\right),\overline{V}\right)d\tau=E_S\left(\overline{W}\right)+\left(\nabla E_S\left(\overline{W}\right),\overline{V}\right)+\int_0^1\left(\nabla E_S\left(\overline{W}+\tau\overline{V}\right)-\nabla E_S\left(\overline{W}\right),\overline{V}\right)d\tau$$

and substituting $\overline{W}=\overline{W}(t)$ and $\overline{V}=-\alpha(t)\nabla E_S\left(\overline{W}(t)\right)$, we obtain:

$$E_S\left(\overline{W}(t+1)\right)=E_S\left(\overline{W}(t)\right)-\alpha(t)\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2-$$

$$-\alpha(t)\int_0^1\left(\nabla E_S\left(\overline{W}(t)-\tau\alpha(t)\nabla E_S\left(\overline{W}(t)\right)\right)-\nabla E_S\left(\overline{W}(t)\right),\nabla E_S\left(\overline{W}(t)\right)\right)d\tau.$$

As the condition (1) is satisfied, then

$$E_S\left(\overline{W}(t+1)\right)\le E_S\left(\overline{W}(t)\right)-\alpha(t)\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2+L\alpha^2(t)\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2\int_0^1\tau d\tau=$$

$$=E_S\left(\overline{W}(t)\right)-\alpha(t)\left(1-\alpha(t)\frac{L}{2}\right)\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2.$$

Considering (3), we have

$$\alpha(t)\left(1-\alpha(t)\frac{L}{2}\right)=\frac{1}{2L}-\frac{L}{2}\left(\alpha(t)-\frac{1}{L}\right)^2\ge\frac{1}{2L}-\frac{L}{2}\left(\varepsilon-\frac{1}{L}\right)^2=\varepsilon-\frac{L}{2}\varepsilon^2=\frac{L}{2}\varepsilon\left(\frac{2}{L}-\varepsilon\right)>0,$$

$E_S\left(\overline{W}(t+1)\right)\le E_S\left(\overline{W}(t)\right)-\frac{L}{2}\varepsilon\left(\frac{2}{L}-\varepsilon\right)\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2$. Summarizing the inequalities $E_S\left(\overline{W}(t+1)\right)\le E_S\left(\overline{W}(t)\right)-\gamma\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2$, where $\gamma=\frac{L}{2}\varepsilon\left(\frac{2}{L}-\varepsilon\right)$, on $t=\overline{0,k}$, we obtain $E_S\left(\overline{W}(k+1)\right)\le E_S\left(\overline{W}(0)\right)-\gamma\sum\limits_{t=0}^{k}\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2$, or

$$\sum_{t=0}^{k}\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2\le\frac{1}{\gamma}\left(E_S\left(\overline{W}(0)\right)-E_S\left(\overline{W}(k+1)\right)\right)\le\frac{1}{\gamma}E_S\left(\overline{W}(0)\right).$$

As $E_S\left(\overline{W}\right)\ge 0$, $\sum\limits_{t=0}^{k}\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2\le\frac{1}{\gamma}E_S\left(\overline{W}(0)\right)$ $\forall k$, i.e. $\sum\limits_{t=0}^{\infty}\left\|\nabla E_S\left(\overline{W}(t)\right)\right\|^2<+\infty$. Hence, the gradient of function $E_S\left(\overline{W}\right)$ aspires to zero: $\lim\limits_{t\to\infty}\nabla E_S\left(\overline{W}(t)\right)=0$, and the function $E_S\left(\overline{W}\right)$ decreases monotonously: $E_S\left(\overline{W}(t+1)\right)\le E_S\left(\overline{W}(t)\right)$. □

Lemma 1. If the gradient of function $E_S(\overline{W})$ satisfies to Lipschitz condition (2), then we obtain: $\left\|\nabla E_S(\overline{W})\right\|^2 \le 2L \cdot E_S(\overline{W})$.

It is obvious, that the result of lemma 1 is carried out for any nonnegative differentiated function $f(\bar{x})$. It means, that $\left\|\nabla f(\bar{x})\right\|^2 \le 2L \cdot f(\bar{x})$. Let consider the function $f(\bar{x}) = \frac{1}{2}(A\bar{x}, \bar{x})$, where a matrix A is symmetric and nonnegative. If we take into consideration, that function $f(\bar{x}) = \frac{1}{2}(A\bar{x}, \bar{x})$ satisfies to conditions of lemma 1 and the equation $\nabla f(\bar{x}) = A\bar{x}$ takes place, i.e. the condition (1) is satisfied with the constant $L = \|A\|$, we obtain $\|A\bar{x}\|^2 = \left\|\nabla f(\bar{x})\right\|^2 \le 2L \cdot f(\bar{x}) = 2\|A\| \cdot \frac{1}{2}(A\bar{x}, \bar{x}) = \|A\| \cdot (A\bar{x}, \bar{x})$. Thus $(A\bar{x}, \bar{x}) \ge \frac{1}{\|A\|}\|A\bar{x}\|^2$ $\forall \bar{x}$ for symmetric and nonnegative matrix A. □

Lemma 2. If $E_S(\overline{W})$ is twice differentiated convex function: $E_S(\lambda\overline{W} + (1-\lambda)\overline{V}) \le \lambda E_S(\overline{W}) + (1-\lambda)E_S(\overline{V})$, for any $\overline{W}, \overline{V}, 0 \le \lambda \le 1$, and its gradient $\nabla E_S(\overline{W})$ satisfies to Lipschitz condition (1), then: $\left(\nabla E_S(\overline{W}) - \nabla E_S(\overline{V}), \overline{W} - \overline{V}\right) \ge \frac{1}{L}\left\|\nabla E_S(\overline{W}) - \nabla E_S(\overline{V})\right\|^2$.

Lemma 3. If function $E_S(\overline{W})$ is strongly convex with constant ℓ: $E_S(\overline{W}) \ge E_S(\overline{V}) + \left(\nabla E_S(\overline{V}), \overline{W} - \overline{V}\right) + \frac{\ell}{2}\left\|\overline{W} - \overline{V}\right\|^2$, and $\overline{W}^*$ is its minimum point, then $\left\|\nabla E_S(\overline{W})\right\|^2 \ge 2\ell\left(E_S(\overline{W}) - E_S(\overline{W}^*)\right)$.

Theorem 2. Let function $E_S(\overline{W})$ be differentiated, and its gradient satisfies to Lipschitz condition (1) and function $E_S(\overline{W})$ is strongly convex with constant ℓ: $E_S(\overline{W}) \ge E_S(\overline{V}) + \left(\nabla E_S(\overline{V}), \overline{W} - \overline{V}\right) + \frac{\ell}{2}\left\|\overline{W} - \overline{V}\right\|^2$. If $\alpha(t)$ satisfies to (3), then the method (2) converges to an unique point of a global minimum $\overline{W}^*$ with a speed of a geometrical progression: $\left\|\overline{W}(t) - \overline{W}^*\right\| \le cq^t$ $0 < q < 1$.

Proof. As all conditions of theorem 1 are satisfied, the following inequality is true: $E_S(\overline{W}(t+1)) \le E_S(\overline{W}(t)) - \alpha(t)\left(1 - \alpha(t)\frac{L}{2}\right)\left\|\nabla E_S(\overline{W}(t))\right\|^2$. Considering, that function $E_S(\overline{W})$ is a strongly convex with a constant ℓ, then $\left\|\nabla E_S(\overline{W})\right\|^2 \ge 2\ell\left(E_S(\overline{W}) - E_S(\overline{W}^*)\right)$, where $\overline{W}^*$ is a point of a global minimum, and

$$E_S\left(\overline{W}(t+1)\right) \le E_S\left(\overline{W}(t)\right) - \ell\alpha(t)\left(2 - L\alpha(t)\right)\left(E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right)\right). \quad (4)$$

Subtracting $E_S\left(\overline{W}^*\right)$ from the left and right parts of the inequality (4), we receive

$$E_S\left(\overline{W}(t+1)\right) - E_S\left(\overline{W}^*\right) \le E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right) - \ell\alpha(t)\left(2 - L\alpha(t)\right)\left(E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right)\right) =$$

$$= \left(1 - \ell\alpha(t)\left(2 - L\alpha(t)\right)\right)\left(E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right)\right) = \left(\ell L\alpha^2(t) - 2\ell\alpha(t) + 1\right)\left(E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right)\right).$$

Designating $q(t) = \ell L\alpha^2(t) - 2\ell\alpha(t) + 1$, we have

$$E_S\left(\overline{W}(t+1)\right) - E_S\left(\overline{W}^*\right) \le q(t)\left(E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right)\right) \le q(t)q(t-1)\left(E_S\left(\overline{W}(t-1)\right) - E_S\left(\overline{W}^*\right)\right) \le$$

$$\le \left(\prod_{i=0}^{t} q(i)\right)\left(E_S\left(\overline{W}(0)\right) - E_S\left(\overline{W}^*\right)\right). \text{ Thus, } E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right) \le \left(\prod_{i=0}^{t-1} q(i)\right)\left(E_S\left(\overline{W}(0)\right) - E_S\left(\overline{W}^*\right)\right).$$

As $\alpha(t)$ satisfies to (3), then

$$q(t) = \ell L\alpha^2(t) - 2\ell\alpha^2(t) + 1 = \ell L\left(\alpha(t) - \frac{1}{L}\right)^2 + 1 - \frac{\ell}{L} \ge q\left(\frac{1}{L}\right) = \frac{L-\ell}{L} \ge 0,$$

and

$$q(t) = \ell L\alpha^2(t) - 2\ell\alpha^2(t) + 1 \le q(\varepsilon t) = q\left(\frac{2}{L} - \varepsilon\right) = \ell L\varepsilon^2 - 2\ell\varepsilon + 1 = 1 - \ell L\varepsilon\left(\frac{2}{L} - \varepsilon\right) < 1.$$

Then $0 \le q(t) \le q_0 = 1 - \ell L\varepsilon\left(\frac{2}{L} - \varepsilon\right) < 1$, and hence $E_S\left(\overline{W}(t)\right) \to E_S\left(\overline{W}^*\right)$ at $t \to \infty$.

As for a strongly convex function with a constant ℓ we have $E_S\left(\overline{W}\right) \ge E_S\left(\overline{V}\right) + \left(\nabla E_S\left(\overline{V}\right), \overline{W} - \overline{V}\right) + \frac{\ell}{2}\left\|\overline{W} - \overline{V}\right\|^2$, then assuming $\overline{V} = \overline{W}^*$ and considering that $\nabla E_S\left(\overline{W}^*\right) = \overline{0}$, we obtain $E_S\left(\overline{W}\right) \ge E_S\left(\overline{W}^*\right) + \frac{\ell}{2}\left\|\overline{W} - \overline{W}^*\right\|^2$ or $\left\|\overline{W} - \overline{W}^*\right\|^2 \le \frac{2}{\ell}\left(E_S\left(\overline{W}\right) - E_S\left(\overline{W}^*\right)\right)$, and finally $\left\|\overline{W}(t) - \overline{W}^*\right\|^2 \le \frac{2}{\ell}\left(E_S\left(\overline{W}(t)\right) - E_S\left(\overline{W}^*\right)\right) \le \frac{2}{\ell}q_0^t\left(E_S\left(\overline{W}(0)\right) - E_S\left(\overline{W}^*\right)\right)$.

Thus, $\left\|\overline{W}(t) - \overline{W}^*\right\| \le cq^t$, where $c = \sqrt{\frac{2}{\ell}}$, $0 \le q = \sqrt{q_0} = \sqrt{1 - \ell L\varepsilon\left(\frac{2}{L} - \varepsilon\right)} < 1$. □

Conclusion

Two conditions presented herein can be applied for a training wide variety of neural network. Exact gradient calculation allows for better convergence. We obtained the results for the given concrete network error function, but the used proof technique can be widespread to other classes of error functions with obtaining local analogues of the theorems.

References

1. Golovko V.A. Neural network: training, organization and application. Book 4 (Neurocomputers and their applications) / Editor A.I. Galushkin. - M.: 2001. - 256 p.
2. D. P. Bertsekas, J. N. Tsitsiklis. Gradient convergence in gradient methods with errors // *SIAM Journal in Optimization. 2000* , Vol. 10, No. 3. P. 627-642.

3. J. Cao, J. Wang. Absolute exponential stability of recurrent neural networks with Lipschitz-continuous activation functions and time delays //Neural Networks.- 2004, Vol.17, N.3. P. 379-390.
4. D. Angeli, B. Ingalls, E. D. sontag, and Y. Wang. **Uniform global asymptotic stability of differential inclusions** // *Journal of Dynamical and Control Systems*. 2004, Vol. 10, No. 3. P. 391–412**.**
5. Gladkij I.I., Golovko V.A., Maxnist L.P. Training of neural networks with use of a method of the steepest descent // *Vestnik of Brest state technical university. Physics, mathematics, chemistry.* - Brest: BSTU, 2001. - № 5 - P. 47-55.
6. **V.** Golovko, N. Manyakov, A. Doudkin. Application of Neural Network Techniques to Chaotic Signal Processing //*Optical Memory and Neural Networks (Information Optics)* : Allerton Press, Inc., 2005. - Vol. 13, № 4. - P. 195-215.