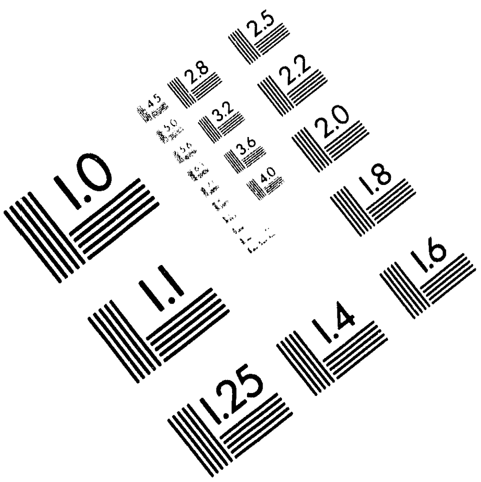
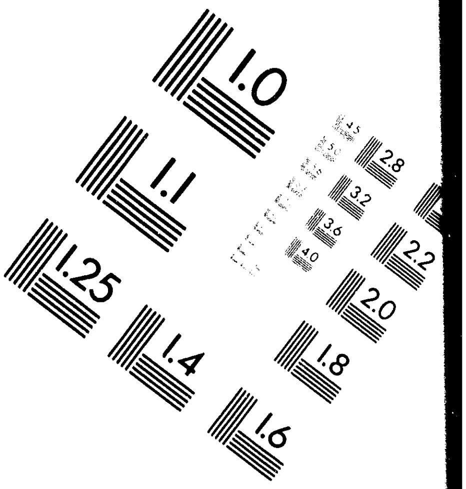
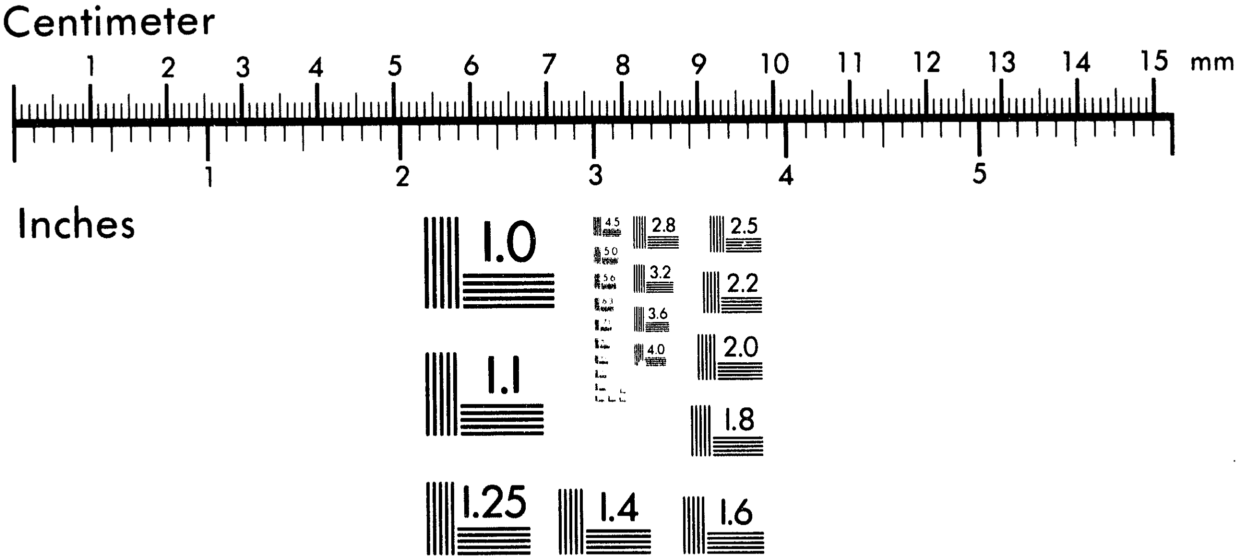
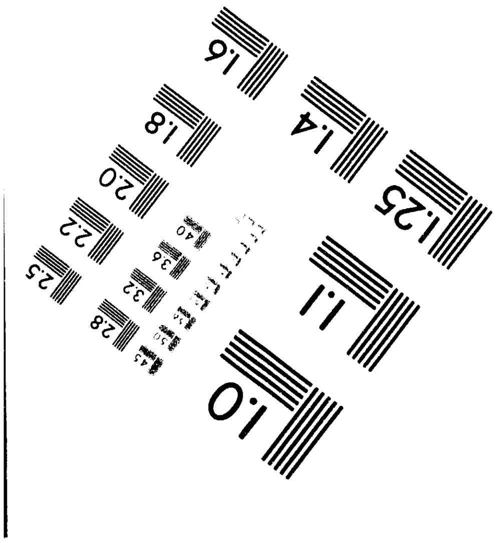
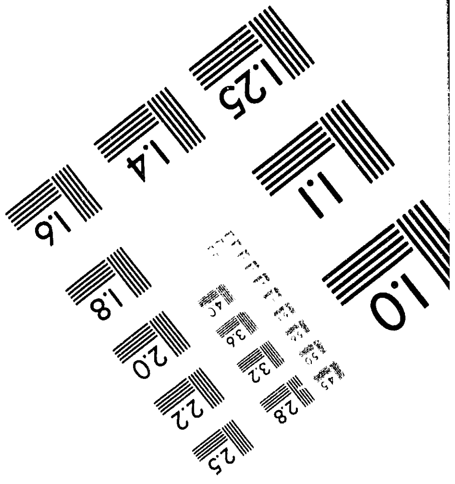





Centimeter
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 mm
1 2 3 4 5
Inches
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3





1 of 1

Paper to be submitted to the IEEE World Congress on Computational Intelligence, June 26-July 2,1994, Orlando, Florida.

The Causes for Premature Saturation with Backpropagation Training*

Javier Vitela[1] and Jaques Reifman[2]

[1]Universidad Nacional Autonoma de Mexico
Instituto de Ciencias Nucleares
04510 Mexico D. F.

[2]Reactor Analysis Division
Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois 60439

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

*Work supported by the U. S. Department of Energy, Nuclear Energy Programs under contract W-31-109-ENG-38.

The Causes for Premature Saturation With Backpropagation Training

Javier E. Vitela

Universidad Nacional Autonoma de Mexico
Instituto de Ciencias Nucleares
04510 Mexico D.F.
vitela@roxanne.nuclecu.unam.mx

Jaques Reifman

Argonne National Laboratory
Reactor Analysis Division
Argonne, Illinois 60439
jreifman@anl.gov

Abstract

The mechanism that causes the output nodes of feedforward multilayer networks mapped with sigmoid functions to prematurely saturate during backpropagation training are described. The necessary conditions for the occurrence of this undesirable phenomenon are also presented. Simulation results demonstrate the adequacy of the presented necessary conditions.

Introduction

The occurrence of the undesirable phenomenon of premature saturation of the output nodes of feedforward multilayer networks mapped with sigmoid functions has been identified[1-5] as one of the causes for decreasing the rate of convergence of the backpropagation (BP) training algorithm. However, the literature related to this phenomenon, which is also known as the flat spot problem, describes only the consequences but not the mechanism that causes premature saturation. The purpose of this paper is to describe the mechanism that causes this undesirable phenomenon and present the necessary conditions for its occurrence.

Backpropagation And Premature Saturation

The BP algorithm trains a feedforward multilayer network by iteratively searching for a set of weights **w** in weight-space that minimize the total training error E over all J_L nodes of the *L-th* or output layer of the network

$$E = \sum_{j=1}^{J_L} E_j = \sum_{j=1}^{J_L} \frac{1}{2} \sum_{p=1}^{P} (t_{pj} - o_{pj}^{(L)})^2. \tag{1}$$

Here, the training error E_j for output node $j=1,2,\ldots,J_L$ is defined as the squared difference between the desired output t_{pj} and the network predicted output $o_{pj}^{(L)}$ over all patterns $p=1,2,\ldots,P$ of the training set. At each iteration k, the weights **w** are updated through the weight-update rule[6]

$$\Delta \mathbf{w}_k = -\eta \nabla E(\mathbf{w}_k) + \alpha \Delta \mathbf{w}_{k-1} = -\eta \sum_{j=1}^{J_L} \nabla E_j(\mathbf{w}_k) + \alpha \Delta \mathbf{w}_{k-1}, \tag{2}$$

where η and α are positive constants smaller than 1.0 known as the learning and momentum parameters, respectively. The component of the gradient ∇E_j corresponding to weight $w_{ni}^{(\ell)}$ connecting the *i-th* node in the *(ℓ-1)-th* layer with the *n-th* node of the *ℓ-th* layer is given by the delta rule

$$\frac{\partial E_j}{\partial w_{ni}^{(\ell)}} = -\sum_{p=1}^{P} \delta_{pn}^{(\ell)}\, o_{pi}^{(\ell-1)}, \tag{3}$$

and similar to the BP algorithm, if the nodes are mapped with a sigmoid function, then

$$\delta_{pn}^{(\ell)} = \begin{cases} (t_{pj} - o_{pj}^{(L)})\, o_{pj}^{(L)}\, (1 - o_{pj}^{(L)}); & \text{for } \ell = L,\ n = j, \\ 0; & \text{for } \ell = L,\ n \neq j, \\ o_{pn}^{(\ell)}\, (1 - o_{pn}^{(\ell)}) \sum_{m=1}^{J_{\ell+1}} \delta_{pm}^{(\ell+1)}\, w_{mn}^{(\ell+1)}; & \text{for } 1 < \ell < L. \end{cases} \tag{4}$$

In the literature, the phenomenon of premature saturation of the network output nodes is characterized when the network predicted output $o_{pj}^{(L)}$ of one or more output nodes approaches values close to either 0 or 1 for all patterns p=1,2,...,P. When that happens, the term $o_{pj}^{(L)}(1 - o_{pj}^{(L)})$in Eq. (4) corresponding to the slope of the sigmoid function approaches zero which, in turn, causes $\partial E_j/\partial w_{ni}^{(\ell)}$ in Eq. (3) to also approach zero. As a consequence, the weights directly connected to the saturated output node j are negligibly updated at each subsequent iteration (until the eventual recovery from saturation) causing them to become *trapped* at their current values. The trapping of the weights may preclude any significant changes in E_j causing an unnecessary decrease in the rate of convergence of the algorithm. The trapping of the weights is generally characterized by regions of flat plateaus in the training curve.

The Mechanism that Causes Premature Saturation

Our analysis shows that premature saturation is likely to occur at the early stages of training when the randomly selected weights place the starting point of the BP algorithm in a region of weight-space that has a skewed error-surface. The skewness of the error-surface may then cause a change in signs at two consecutive iterations in the components of $\nabla E_j(\mathbf{w}_k)$ directly connected to output node j. This may cause the projection of $\alpha\Delta\mathbf{w}_{k-1}$ in Eq. (2) along the direction of ∇E_j not to point in the desired negative gradient direction $-\nabla E_j$. If in addition, the magnitude of the projection of $\alpha\Delta\mathbf{w}_{k-1}$ along ∇E_j is larger than the corresponding magnitude of $-\eta\nabla E(\mathbf{w}_k)$, the effect, to first order, of the new weight update $\Delta\mathbf{w}_k$ will be an increase in E_j.

At the early stages of the training process, the components of $\Delta\mathbf{w}_k$ may be of the same order of magnitude as the components of the weight $\mathbf{w}_k$. Therefore, the updated weight $\mathbf{w}_{k+1} = \mathbf{w}_k + \Delta\mathbf{w}_k$ may suddenly cause the network predicted output $o_{pj}^{(L)}$ of one or more of the output nodes to approach either 0 or 1, for all patterns p=1,2,...,P, in the following iteration. The combination of the effects of this scenario, for output node j, with the effects of a skewed error-surface would cause an increase in $E_j(\mathbf{w}_{k+1})$, a decrease in the magnitude of $\nabla E_j(\mathbf{w}_{k+1})$, and a reduction in the contribution of $\nabla E_j(\mathbf{w}_{k+1})$ to the total gradient $\nabla E(\mathbf{w}_{k+1})$. If the updated learning term $-\eta\nabla E(\mathbf{w}_{k+1})$ cannot offset the tendency of the momentum term $\alpha\Delta\mathbf{w}_k$ to update the weights in the direction in which E_j increases, the magnitude of ∇E_j will be further reduced in the following iteration. As this undesirable mechanism persists for consecutive iterations, the components of ∇E_j approach zero strongly affecting the weights directly connected to output node j. As ∇E_j approaches zero and the contribution of the momentum term for these weights decrease at consecutive iterations (since α is smaller than 1.0), the weights directly connected to node j become trapped at their current values until the eventual recovery from saturation.

In general, premature saturation is not manifested in all output nodes of the network. The occurrence of this phenomenon as well as the number of saturated nodes are strongly dependent on the starting point in weight-space, the values of η and α, the topology of the network, and the size and type of the training patterns. These dependencies together with the strong nonlinearity of sigmoid-mapped nodes offer tremendous difficulties in obtaining both the necessary and sufficient conditions for the occurrence of premature saturation. In the following, we present only a set of *necessary* conditions that need to be satisfied if an output node is to saturate prematurely.

Necessary Conditions for Premature Saturation

To first order, each iteration of the BP algorithm yields smaller values for E_j if the weight-update $\Delta \mathbf{w}_k$ satisfies the inequality $\Delta \mathbf{w}_k \cdot \nabla E_j(\mathbf{w}_k) < 0$ for output node j. Based on the fact that this inequality is not satisfied at the onset of premature saturation, i.e., E_j increases, we may introduce the following four necessary conditions that must be satisfied at several consecutive iterations at the early stages of training if premature saturation is to occur:

(c1) $\eta \nabla E(\mathbf{w}_k) \cdot \nabla E_j(\mathbf{w}_k) - \alpha \Delta \mathbf{w}_{k-1} \cdot \nabla E_j(\mathbf{w}_k) < 0$,

(c2) $\Delta \mathbf{w}_{k-1} \cdot \nabla E_j(\mathbf{w}_k) > 0$,

(c3) $\alpha \, |\Delta \mathbf{w}_{k-1} \cdot \nabla E_j(\mathbf{w}_k)| >> \eta \, |\nabla E(\mathbf{w}_k) \cdot \nabla E_j(\mathbf{w}_k)|$, and

(c4) $|\nabla E_j(\mathbf{w}_{k+1})| < |\nabla E_j(\mathbf{w}_k)|$.

These four conditions basically summarize the mechanism that causes premature saturation. Condition c1 is obtained by substituting $\Delta \mathbf{w}_k$ of Eq. (2) in the inequality $\Delta \mathbf{w}_k \cdot \nabla E_j(\mathbf{w}_k)<0$ and expresses the fact that E_j should increase at the onset of premature saturation. Condition c2 implies that the projection of $\Delta \mathbf{w}_{k-1}$ along the direction of $\nabla E_j(\mathbf{w}_k)$ should point in the incorrect or gradient direction, as opposed to the correct or negative gradient direction. Condition c3 is satisfied if the magnitude of the projection of the momentum term is larger than the projection of learning term along $\nabla E_j(\mathbf{w}_k)$. This is necessary if the momentum term $\alpha \Delta \mathbf{w}_{k-1}$ is to govern the motion of the weights across the error-surface. Finally, condition c4 expresses the fact that the magnitude of the gradient of node j should decrease at consecutive iterations reflecting the fact that node j is becoming saturated. The magnitude of the gradient ∇E_j can then be used as a measure of the degree of saturation of node j.

Although necessary, these conditions are not sufficient for premature saturation to occur. Even if these conditions are satisfied for a given output node during consecutive iterations, the node may only show a slight tendency to saturate. That is due to the fact that the occurrence of premature saturation is an overall property of the network with the output nodes that remain unsaturated playing important roles. For example, in some training sessions the contribution of the unsaturated nodes to the weight update $\Delta \mathbf{w}_k$ may prevent the saturated nodes from reaching large degrees of saturation at early stages of the saturation process. In this case, the saturated nodes will recover quickly. On the other hand, in other training sessions the contribution of the unsaturated nodes may inhibit a faster recovery of the saturated nodes at later stages of the saturation process.

Simulation Results

To illustrate the above discussions, we present the results of a BP training session for the classification of three component failures in a nuclear power plant[5] in which the network output nodes saturate prematurely. The network consisted of three layers with 20-20-3 nodes per layer, respectively, and the desired output values t_{pj} for the three output nodes were set to either 0.1 or 0.9 depending on the training pattern. The value of the learning parameter η was fixed at 0.1 throughout the training session and the value of the momentum parameter α was set to 0.0 for the first two training cycles and after that it was set to 0.9. A training cycle or an iteration in the BP algorithm consisted of the presentation of the entire set of 108 training patterns, with 36 patterns for each one of the three component failures, after which, the weights were adjusted.

Figures 1 and 2 show the training results for the case in which a set of randomly selected weights caused two nodes, 1 and 3, out of the three output nodes to saturate prematurely. The occurrence of saturation for these two nodes is characterized in Fig. 1 by the two overlapping plateaus in the training curves for nodes 1 and 3. The formation of the two plateaus is the consequence of the trapping of the weights associated with the two nodes. The plateaus remain until the eventual recovery of the nodes at around 350 and 28,000 training cycles, respectively, for nodes 3 and 1.

Premature saturation starts at the third training cycle and is represented in Fig. 1 by a slight increase in the values of the training errors E_1 and E_3 which satisfies necessary condition c1. The increase in the values of the training errors is caused by incorrect updates of the weights directly connected to these two nodes due to the satisfaction of necessary conditions c2 and c3. The momentum term $\alpha\Delta w_{k-1}$ governs the incorrect motion of the weights across the error-surface. Figure 2 illustrates that starting at the third training cycle the magnitude of the gradients ∇E_1 and ∇E_3 decrease at consecutive iterations (until the nodes become completely saturated around 20 cycles) satisfying necessary condition c4. In this training session, each one of the four necessary conditions was satisfied for nodes 1 and 3 between iterations 3 and 14, after which, the nodes remained saturated for a number of iterations until the eventual recovery from saturation.

Figure 2 also illustrates the degree of saturation and the recovery of the two nodes from saturation. The values of $|\nabla E_1|$ and $|\nabla E_3|$ express the degree of saturation of output nodes 1 and 3, respectively, which play an important role in their recovery from saturation. The smaller the magnitude of the gradient ∇E_j is at the last cycle for which the four necessary conditions are satisfied, the longer is the recovery process from saturation. The value of $|\nabla E_3|$ is larger than the value of $|\nabla E_1|$ at iteration 14 allowing for a faster recover of node 3. The values of $|\nabla E_j|$ increase during the recovery process allowing the weights directly connected to output node j to come out of their trapped state. Thereafter, E_j decreases significantly as the algorithm resumes its motion towards the optimum solution.

Summary and Conclusions

Additional experiments were performed in order to confirm the description of the underlying mechanism and necessary conditions for the occurrence of premature saturation of network output nodes during BP training. The experiments included changing the values of the learning and momentum parameters as well as running the same problem with different sets of initial weights. All our experiments confirmed the results presented here.

Acknowledgments

The second author was supported by the U.S. Department of Energy, Nuclear Energy Program, under contract number W-31-109-ENG-38.

References

1. S. E. Fahlman, "Faster-Learning Variation on Back-Propagation: An Empirical Study," *Proc. of the 1988 Connectionist Models Summer School*, Carnegie-Mellon University, D. Touretzky, G. Hinton, and T. Sejnowski, Eds., M. Kaufmann, San Mateo, CA, 1988.

2. Y. Lee, S. H. Oh, and M. W. Kim, "The Effect of Initial Weights on Premature Saturation in Back-Propagation Learning," *IJCNN*, Seattle, Washington, July 8-12, 1991.

3. R. Parekh, K. Balakrishnan, and V. Honavar, "An Empirical Comparison of Flat-Spot Elimination Techniques in Back-Propagation Networks," *Proc. Third Workshop on Neural Networks*, Auburn, Alabama, February 10-12, 1992.

4. J. E. Vitela and J. Reifman, "Enhanced Backpropagation Training Algorithm for Transient Event Identification," *Trans. Am. Nucl. Soc.*, **69**, 148, 1993.

5. J. Reifman and J. E. Vitela, "Accelerating Learning of Neural Networks with Conjugate Gradients for Nuclear Power Plant Applications," to appear in *Nuclear Technology*, May, 1994.

6. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. I, D. E. Rumelhart and J. C. McClelland, Eds., MIT Press, 1986.

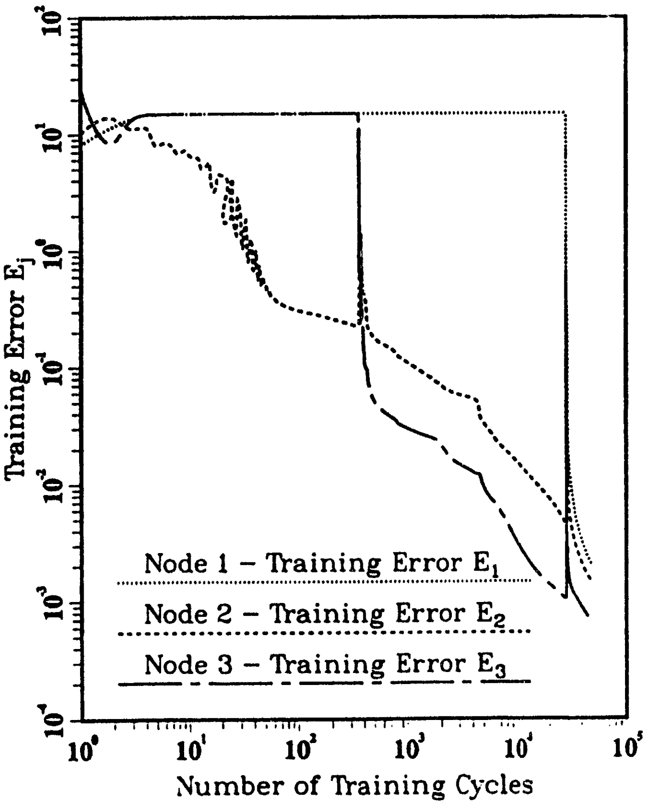


Fig. 1. Effects of the Occurrence of Premature Saturation of Two Output Nodes on the Training Errors of a Feedforward Three-Layer Network

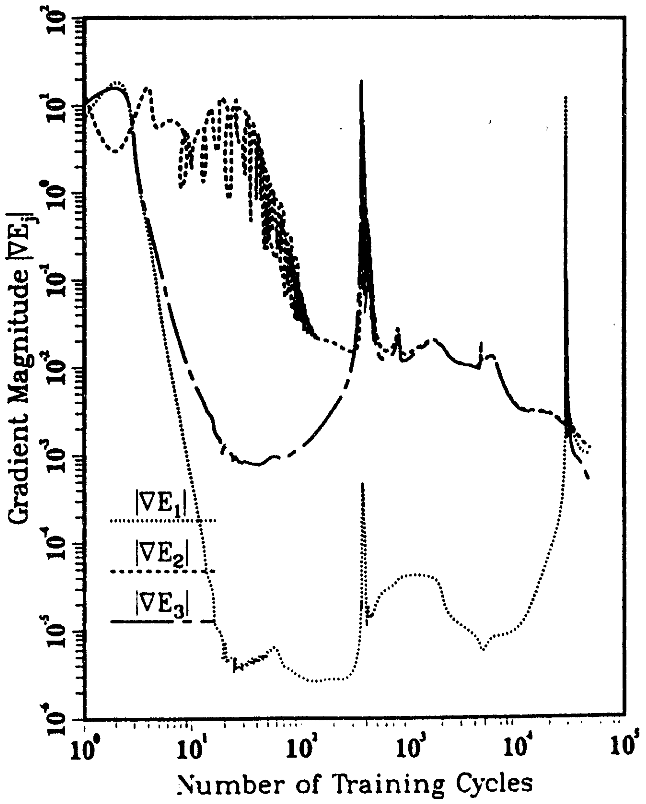


Fig. 2. Behavior of the Magnitude of the Gradient of the Training Errors During a Training Session in which Premature Saturation Is Observed

DATE
FILMED
7/25/94

END